Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2020"

Moscow, June 17–20, 2020

RUREBUS-2020 SHARED TASK: RUSSIAN RELATION EXTRACTION FOR BUSINESS

Ivanin V. A. (vitalii.ivanin@abbyy.com)[1, 2],
Artemova E. L. (Echernyak@hse.ru)[3],
Batura T. V. (tatiana.v.batura@gmail.com)[4, 7],
Ivanov V. V. (nomemm@gmail.com)[5, 7],
Sarkisyan V. V. (vsarkisyan@hse.ru)[3],
Tutubalina E. V. (elvtutubalina@kpfu.ru)[6, 7],
Smurov I. M. (ivan.smurov@abbyy.com)[1, 2]

[1]ABBYY, [2]Moscow Institute of Physics and Technology,
[3]National Research University Higher School of Economics,
[4]Novosibirsk State University, [5]Innopolis University,
[6]Kazan Federal University,
[7]Lomonosov Moscow State University

In this paper, we present a shared task on core information extraction problems, named entity recognition and relation extraction. In contrast to popular shared tasks on related problems, we try to move away from strictly academic rigor and rather model a business case. As a source for textual data we choose the corpus of Russian strategic documents, which we annotated according to our own annotation scheme. To speed up the annotation process, we exploit various active learning techniques. In total we ended up with more than two hundred annotated documents. Thus we managed to create a high-quality data set in short time. The shared task consisted of three tracks, devoted to 1) named entity recognition, 2) relation extraction and 3) joint named entity recognition and relation extraction. We provided with the annotated texts as well as a set of unannotated texts, which could of been used in any way to improve solutions. In the paper we overview and compare solutions, submitted by the shared task participants. We release both raw and annotated corpora along with annotation guidelines, evaluation scripts and results at https://github.com/dialogue-evaluation/RuREBus.

Keywords: named entity recognition, relation extraction, shared task, Russian fine-tuning, BERT

DOI: 10.28995/2075-7182-2020-19-416-431

RUREBUS-2020: СОРЕВНОВАНИЕ ПО ИЗВЛЕЧЕНИЮ ОТНОШЕНИЙ В БИЗНЕС-ПОСТАНОВКЕ

Иванин В. А. (vitalii.ivanin@abbyy.com)[1, 2],
Артемова Е. Л. (echernyak@hse.ru)[3],
Батура Т. В. (tatiana.v.batura@gmail.com)[4, 7],
Иванов В. В. (nomemm@gmail.com)[5, 7],
Саркисян В. В. (vsarkisyan@hse.ru)[3],
Тутубалина Е. В. (elvtutubalina@kpfu.ru)[6, 7],
Смуров И. М. (ivan.smurov@abbyy.com)[1, 2]

[1]ABBYY, [2]Московский Физико-технический Институт
[3]Национальный исследовательский университет Высшая школа экономики
[4]Новосибирский государственный университет
[5]Иннополис, [6]Казанский федеральный университет
[7]МГУ им. М. В. Ломоносова

В статье представлены результаты соревнования по распознаванию именованных сущностей и извлечению отношений. Целью соревнования является сравнение методов извлечения сущностей и отношений на русском языке в постановке, приближенной к индустриальным задачам. В качестве исходной коллекции текстов использовался корпус Минэкономразвития РФ, содержащий программы стратегического развития. Корпус был размечен в соответствии с инструкцией, разработанной авторами статьи. В процессе разметки использовались различные методы активного обучения, что позволило за короткое время создать качественный набор данных. Всего было размечено более двухсот документов. Соревнование проводилось по трем задачам (дорожкам): 1) распознавание именованных сущностей, 2) извлечение отношений и 3) совместное распознавание именованных сущностей и извлечение отношений. Вместе с коллекцией размеченных текстов участникам также были предоставлены неразмеченные тексты, которые могли быть использованы для улучшения решений. В статье дается обзор и сравниваются результаты участников соревнования. Детальное описание соревнования, текстовые коллекции, инструкция по разметке и скрипты для оценки качества доступны по ссылке: https://github.com/dialogue-evaluation/RuREBus.

Ключевые слова: распознавание именованных сущностей, извлечение отношений, соревнование, русский, дообучение, BERT

1. Introduction

Structuring unstructured information is one of the most popular industrial application of natural language processing. Standard approaches to it require named entity recognition (NER) and/or relation extraction (RE). NER and RE are classical NLP tasks, formulated as early as mid-1990s [39]. There exist a number of well-studied academic corpora (see next section for multiple examples of such corpora). Scores obtained on these corpora are typically high. Taking recent advances in NER in account one can even assume that it is a largely solved task.

However, business applications seldom do enjoy the high scores reported in academia. In our opinion the main reason for that is the fact that both text sources and entities in industry and academia present with several noticeable differences.

Firstly, business case texts are usually domain-specific (e. g. legal) texts that can contain less than perfect language or other irregularities (ponderous sentences with complicated syntactic structure, slang etc.). Academic baselines, on the other hand, typically consist of well-written news or biography (or scientific in case of BioNLP) texts without any irregularities of this kind.

Secondly, while entities in academia are usually compact and well-defined, industry sometimes has to deal with something much more loose, spanning for many words and with less than clear borders.

Our main motivation for conducting this work was to attempt to bridge the gap between academic NLP and less-than-ideal business scenarios. In order to do so, we have collected and marked up a corpus of governmental documents, produced by the Ministry of Economic Development of the Russian Federation and organized a shared task on it, which are two main contributions of this paper.

2. Related work

In this section we number related research areas:

1. general domain named entity recognition and relation extraction
2. methods for named entity recognition and relation extraction
3. named entity recognition and relation extraction for the Russian language

2.1. General domain named entity recognition and relation extraction

The entity recognition task is a necessary stage of extracting information from texts. Today there are quite a lot of datasets for the task in different languages for the general domain, such as CONLL 2003 [39], MUC-6 [17], OntoNotes 5 [19], etc.

These datasets usually poses a few types of named entities, such as *persons*, *organizations*, *locations* and casual relations, such as *being born in*, *have position at*, etc.

To perform semantic analysis, it is also important to extract relations that link named entities. This requires building datasets for solving the problem of relation extraction. The relation extraction problem goes further than named entity recognition, as it requires greater understanding of language semantics.

Therefore, there are fewer datasets available both for named entity recognition and relation extraction. The most used datasets are CONLL04 [10], ACE 2005 [47], TACRED [44], SemEval 2010 Task 8 [18].

2.2. Methods for named entity recognition and relation extraction

At the core of the majority of current methods both for named entity recognition and relation extraction are pre-trained language models, such as ELMo [31], BERT [14] and their descents. Using pre-trained language models does not require training a model from scratch, but rather fine-tuning the model for the task under consideration. An example of BERT fine-tuning is presented in [37]. To achieve relation representation by fine-tuning BERT with a large scale "matching the blanks" pretraining entity linked texts are used. This method performs well on the SemEval 2010 Task 8 dataset (F1-measure of 89.5%) and outperforms previous methods on TACRED (F1-measure of 71.5%). For the entity recognition task, the BERT-MRC model [25] achieves the best results on ACE 2005 (F1 score of 86.88%).

The state-of-the-art approach to relation extraction is an entity pair graph-based neural network (EPGNN) model, relying on a graph convolutional network [45]. EPGNN combines sentence semantic features generated by a pre-trained BERT model with graph topological features for relation classification. It shows a macro-F1 score of 90.2% on the SemEval 2010 Task 8 dataset and a micro-F1 score of 77.1% on ACE 2005.

2.3. Named entity recognition and relation extraction for the Russian language

To the best of our knowledge, several datasets for named entity recognition in the Russian language are available: the dataset, developed by Gareev et al. [16], Persons 1000 and Collection 5 [30], [40], [42], FactRuEval 2016 [7], the Russian subset of the BSNLP Shared Task [33].

Prior to deep and even machine learning methods, rule-based approaches dominated the information extraction systems [11], [15]. Most of the early works for the Russian language NER describe systems based on linguistic resources: dictionaries, templates, and rules [9], [12], [35]. Popov et al. described the adaptation of the vocabulary approach for the Russian language [35]. Craidlin introduced the TagLite program, which aims to distinguish named groups consist of three types of proper names: persons, organizations and geographical objects [12]. The system includes the following dictionaries: proper names, generic concepts of investigated entity types (director, river, office) and other auxiliary words that can be part of target noun groups. In order to resolve the ambiguity and process words that are not encountered in dictionaries, the rule-based "predictor" module is applied. The authors evaluated the quality of the system on their own annotated corpus. TagLite obtained 85.8% of F-measure for all categories of named groups. Brykina et al. proposed a system that recognizes named entities based on lists of terms from the input ontology and resolving polysemy with a set of manually developed rules and dictionaries of context words [9]. The authors evaluated the efficiency of a system on their own corpus, considering only entities included in the ontology. The system obtained F-measure varying from 91% to 98% for different types of entities. Both systems were evaluated on closed corpora, which makes it difficult to conduct a comparative analysis of the achievements in this area.

Starting from 2013, studies about Russian NER [4, 16, 34] started to apply Conditional Random Fields (CRF) [23]. Antonova et al. applied a CRF model to their own annotated corpora consisting of news feed texts [4]. There were five types of annotations: person names, geographical objects, organization, products, and events. The authors also evaluated different types of optimizers for CRF. The highest F-measure obtained by this approach was 87.18%. Podobryaev applied CRF model to person recognition and used information from ontology as one of the features [34]. The quality of the proposed approach was evaluated on a manually annotated corpora. Gareev et al. developed an annotated corpus of Russian-language texts for evaluating NER methods and compared the effectiveness of two approaches [16]. The first approach is based on dictionaries of names and rules, which analyze the context of a named entity and compare the set of references to the same entity in a document. The second is based on the CRF model with various features. The developed corpus is publicly available and contains two types of annotations: persons and organizations. The results of experiments showed that CRF-based approach outperformed knowledge-based approach on 13% of F-measure. Mozharova and Loukachevitch investigated the knowledge and context features for the CRF model in the NER task [30].

Recent works on Russian NER focused on neural network models. Anh et al. [2] investigated a combination of bidirectional Long Short Term Memory (LSTM) with CRF and word embeddings presented in [20], [24], [26]. This model showed the best results on three Russian language data sets Gareev's [16], Person-1000 [40] and FactRuEval 2016 [7]. In the work [28], authors showed close to the state of the art results at the time, exploiting only neural model trained on the small dataset without pre-training. There is another aspect of robustness of LSTM-CRF in NER task addressed in the paper [27]. The experiments were conducted in three datasets, Persons-1000 for Russian language, CAp'2017 for French, and CoNLL'03 for English. Remarkably a proprietary system, which combines rule-base approach to statistics analysis, achieves state of the art [7] for the FactRuEval dataset. For fair comparison it should be noted, that proprietary systems are being developed for longer period of time, while the majority of shared task participants train the models from scratch without possessing rule and code base.

The BSNLP 2019 Shared Task [32] introduced a new multilingual dataset, annotated with named entities for four Slavic languages, Bulgarian, Czech, Polish, Russian. The named entities considered were persons, locations, organizations, events, products. The majority of the systems which participated in the shared tag exploited BERT-based solutions by either fine-tuning multilingual BERT [41] or by training the BERT model from scratch for the target languages [5].

Much less attention is paid to relation extraction for the Russian language compared to named entity recognition. To the best of our knowledge, three datasets are annotated with relations. The aforementioned FactRuEval dataset [7] provides with two layers of annotations, the first layer being named entities and the second layer being relations between them. As FactRuEval covers news domain the relations annotated express attribution and occupation properties as well as some facts, such as meetings and deals. Existing datasets [22], [42] are much smaller and are not widely used for experiments.

3. Data

In the RuREBus Shared Task, we proposed a corpus with annotations of named entities and relations. The novelty of the task is in its focus on methods for the extraction of business-relevant entities and relations from corporate documents. For obvious reasons, it is hard to find such documents online and make a shared collection of business-related documents.Therefore, we use a similar collection of corporate documents shared by the Ministry of Economic Development of Russia.

3.1. General corpus description

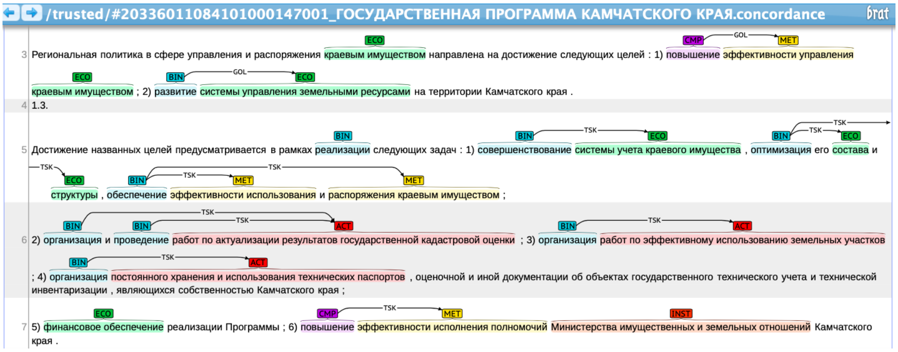
The collection shared by the Ministry of Economic Development of Russia contains strategic planning programs of development for Russian regions. The corpus was studied in [1]. It has 85,501 documents, 298,809,024 tokens overall. Key features of the corpus include:

- uniformity of texts: documents have the same domain, purpose, very similar style and size;
- shared scope: documents mention various types of economic and social entities and relations at different levels of management;
- fixed modalities: a fixed list of modalities in documents that cover current state of the economy or society (problems), as well as plans for future (actions, tasks, etc.)

For the purpose of the RuREBus Task we selected and annotated 218 documents. The annotation guidelines and results of manual annotation are presented below and available at https://github.com/dialogue-evaluation/RuREBus.

3.2. Annotation Guidelines

To support consistency of markup we developed an instruction for entity and relation identification. We also use Brat Rapid Annotation Tool (BRAT) [38] to provide an annotation interface for assigning entities and relations. Preview may be seen in Figure 1.



/trusted/#20336011084101000147001_ГОСУДАРСТВЕННАЯ ПРОГРАММА КАМЧАТСКОГО КРАЯ.concordance

3 Региональная политика в сфере управления и распоряжения краевым имуществом направлена на достижение следующих целей : 1) повышение эффективности управления краевым имуществом ; 2) развитие системы управления земельными ресурсами на территории Камчатского края .

4 1.3.

5 Достижение названных целей предусматривается в рамках реализации следующих задач : 1) совершенствование системы учета краевого имущества , оптимизация его состава и структуры , обеспечение эффективности использования и распоряжения краевым имуществом ;

6 2) организация и проведение работ по актуализации результатов государственной кадастровой оценки ; 3) организация работ по эффективному использованию земельных участков ; 4) организация постоянного хранения и использования технических паспортов , оценочной и иной документации об объектах государственного технического учета и технической инвентаризации , являющихся собственностью Камчатского края ;

7 5) финансовое обеспечение реализации Программы ; 6) повышение эффективности исполнения полномочий Министерства имущественных и земельных отношений Камчатского края .

Figure 1: Annotation interface for assigning entities and relations

We partly employ double annotation. For each annotator we compare several documents with another independent annotation by the verified annotator. This mirroring helps the moderator to resolve arguable cases. After moderation we consider annotator as experienced enough and approve markup without doubling. However, we moderate each document manually even for experienced annotators.

The Cohen's kappa measured on the documents that were marked up twice (not taking into account moderators) is equal to 0.698.

3.3. Entity Descriptions

We developed eight entity types for annotation. Entities are described in Table 1.

Table 1: Entity descriptions

Entity	Entity description	Examples (English)	Examples (Russian)
MET (metric)	indicator or object on which the comparison operation is defined	students' education level total length of roads birth rate economic growth	уровень образования студентов общая протяженность дорог уровень рождаемости экономический рост
ECO (economics)	economic entity (excluding MET) or infrastructure object	private business PJSC Gazprom fuel and energy complex library and museum funds	частный бизнес ПАО Газпром топливно-энергетический комплекс библиотечные и музейные фонды
INST (institution)	institutions, structures and organizations	Youth Employment Center Family and Child Support Organizations metro stations road system	Центр занятости молодёжи Организации поддержки семьи и детства станции метрополитена система дорог
BIN (binary)	binary characteristics or single action	modernization rendering is functioning absence	модернизация оказание функционирует отсутствие
CMP (compare)	comparative characteristic	increase saturation excess of level negative dynamics	рост насыщение превышение уровня негативная динамика
QUA (qualitative)	quality characteristic	effective stable safe poorly developed	эффективный стабильный безопасный плохо развитый
SOC (social)	social object	scientific and educational potential leisure activities folk art the youth	научный и образовательный потенциал досуг народное творчество молодежь
ACT (activity)	activities, events or measures taken by the authorities; these entities are often combined with BIN, e.g., <BIN> developed </BIN> <ACT>an educational project for rural schools </ACT>	restoration work educational project "University 2020" orphan prevention weekend fair	реставрационные работы образовательный проект «Университет 2020» профилактика сиротства ярмарка выходного дня

3.4. Relations Description

We define two relations to describe plans and goals, and nine to describe state of affairs. These relation types could be useful in specific practical applications [6]. **GOL** relation represents abstract goals and aims of the program, e.g., *birth rate increase*. These goals are some objectives that must be achieved as a result of programs' actions.

TSK relation corresponds to concrete tasks and actions taken to achieve some goals, e.g., *opening of new metro stations*.

The other nine relations can be grouped by two criteria: time component (past **P**, present **N**, future **F**) and estimation component (negative **NG**, neutral **NT**, positive **PS**). Past negative, neutral and positive relations (**PNG**, **PNT**, **PPS** respectively) denote implemented changes, present relations (**NNG**, **NNT**, **NPS**) describe the current state of affair, and future relations (**FNG**, **FNT**, **FPS**) present plans and forecasts.

The examples of annotated relations are shown in Table 2.

Table 2: Relation examples

Relation	Example (English)	Example (Russian)
GOL	<CMP> increasing </CMP> <MET> accessibility of transport services </MET>	<CMP> увеличение </CMP> <MET> доступности транспортных сервисов </MET>
TSK	hospital <ACT> overhaul </ACT>	<ACT> капитальный ремонт </ACT> больницы
PPS	<ACT> road works </ACT> <BIN> are completed </BIN>	<ACT> дорожные работы </ACT> <BIN> завершены </BIN>
FNG	<ECO> ruble exchange rate </ECO> <CMP> is expected to drop </CMP>	<CMP> ожидается снижение </CMP> <ECO> курса обмена рубля </ECO>
NNT	<ECO> salary level </ECO> <BIN> stabilized </BIN>	<ECO> уровень заработной платы </ECO> <BIN> стабилизировался </BIN>

3.5. Active learning

We also use an active learning technique [36] to help the annotators and speed up their work. Firstly we obtained a subset of the corpus marked with defined named entities and relations. Next we trained the NER model and employ it further to markup unlabeled documents. Then documents were marked up by annotators. The annotations were verified by moderators. After obtaining new parts of the final corpus model were retrained with this part added to training set. Full pipeline could be seen in Figure 2.

In this work we employ a basic NER model, namely char-CNN-BiLSTM-CRF (proposed by Lample et al [24] and further developed by Ma and Hovy [26]). This architecture is widely used as a robust baseline in sequence tagging tasks. We use FastText [8] embeddings trained by RusVectores [21]. We also employ morphological, syntactical and semantical features, obtained from Compreno [3], [46] and some hand-made features, such as capitalization templates and dependency tree distance between relation members.

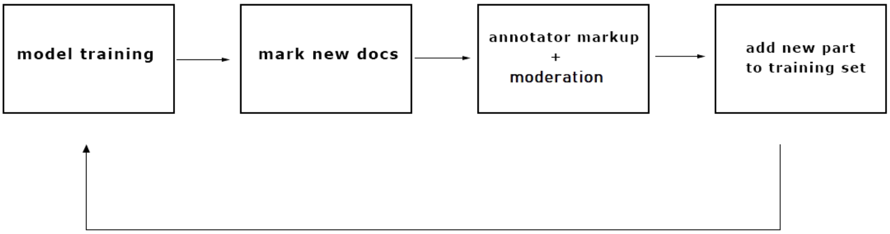


Figure 2: Active learning pipeline

3.6. Basic statistics

Table 3: Statistics of annotated entities

	total	mean len (std)
BIN	30,201	1.05 (0.28)
MET	14,161	4.23 (3.50)
QUA	7,719	1.14 (0.52)
CMP	9,288	1.16 (0.78)
SOC	10,834	2.77 (2.31)
INST	7,903	3.69 (2.81)
ECO	24,853	2.78 (2.19)
ACT	12,274	4.74 (4.57)

We computed descriptive statistics based on annotated documents. Each document was divided into parts by 150 sentences cutoff. In the training set there are 188 annotated parts, where the average number of named entities in a file is 289 and there are 67 relations in average. The mean file length is 1,787 tokens. In the test set there are 30 files, 287 entities and 67 relations on average, the mean length is 1,967 tokens. Tokenization was performed by razdel tokenizer.[1]

4. Shared Task Set-Up

The participants were offered 3 different NER and RE tasks:

1. Named entity recognition.
2. Relation extraction. In this task, the participants were provided with manually annotated named entities. The task was to extract relations between them.
3. End-to-end relation extraction. The participants were to extract both named entities and the relations between them.

[1] https://github.com/natasha/razdel

All three tasks were evaluated with micro-averaged F-measure (evaluation script is available at https://github.com/dialogue-evaluation/RuREBus).

Since Task 2 requires gold standard NER labels, evaluation was organized in two phases. During phase one participants had raw texts of the test set without any markup and were able to solve Tasks 1 and 3. After phase one completion, gold standard labels for all test set texts were provided and evaluation on Task 2 commenced.

During both phases "true" test set was mixed within 514 unannotated texts in order to deny participants the possibility of identifying the exact texts used for evaluation. For phase two, NER markup for these additional texts was obtained with the model used for active learning.

5. Results and analysis

5.1. Results

We have received several submissions after the Shared Task baseline (but before the gold-standard test markup was published). While these results are not considered being part of official Shared Task evaluation, it is prudent to provide this numbers. Participant *davletov-aa* was able to achieve f-measure of 0.132 on Task 3, while *bondarenko* got 0.498 on Task 1.

Table 4: Results of the competition (Micro F1-score). Table is sorted by scores on the NER task, but all 3 tasks are equally important.

Team	NER	RE with NEs	End-to-end RE
davletov-aa	**.561**	.394	—
Sdernal	.464	**.441**	—
ksmith	.463	.152	**.062**
viby	.417	.218	—
dimsolo	.400	—	—
bond005	.338	.045	—
Student2020	.253	—	—

Table 5: F1-score performance measure on the NER task by NE class

Team	ACT	BIN	CMP	ECO	INST	MET	QUA	SOC
davletov-aa	**0.33**	**0.62**	0.79	**0.52**	**0.52**	**0.57**	**0.56**	**0.45**
Sdernal	0.22	0.60	0.77	0.41	0.34	0.43	0.49	0.29
ksmith	0.13	0.52	**0.82**	0.36	0.34	0.41	0.54	0.33
viby	0.21	0.49	0.77	0.29	0.33	0.39	0.52	0.23
dimsolo	0.12	0.50	0.70	0.30	0.37	0.31	0.40	0.29
bond005	0.08	0.56	0.74	0.28	0.16	0.28	0.39	0.14
Student2020	0.10	0.20	0.28	0.28	0.31	0.34	0.34	0.22
average	0.17	0.50	0.70	0.35	0.34	0.39	0.46	0.28

Table 6: F1-score performance measure on the RE task by RE class

Team	NNG	NNT	NPS	FNG	FNT	FPS	PNG	PNT	PPS	GOL	TSK
bond005	0.02	0.02	0.02	0.01	0.01	0.02	0.02	0.01	0.03	0.08	0.14
davletov-aa	**0.63**	0.16	**0.28**	0.34	**0.23**	**0.45**	**0.52**	**0.30**	0.48	0.32	0.42
ksmith	0.25	0.00	0.13	0.05	0.00	0.11	0.00	0.00	0.07	0.27	0.13
Sdernal	0.62	**0.19**	0.24	**0.44**	0.00	0.39	0.23	0.10	**0.58**	**0.43**	**0.47**
viby	0.43	0.01	0.05	0.06	0.05	0.08	0.06	0.07	0.12	0.29	0.23
average	0.39	0.08	0.14	0.18	0.06	0.21	0.16	0.10	0.26	0.28	0.28

One can easily notice that the scores obtained for all tasks are incomparable to the ones usually reported on most widespread academic corpora such as CoNLL-03. In our opinion this fact cannot be attributed to the methods used by participants since most popular approaches were tested by them (as is shown in the next section). However, these less than perfect scores are much closer to the scores often obtained in industry.

Another comparison we can draw is with SemEval-2020 Task 11 [13], a shared task on detecting propaganda spans from text (where various linguistic structures were considered propaganda, such as “Loaded Language”, “Whataboutism, Straw Men, Red Herring” or “Flag-Waving”). Both Propaganda Detection and RuREBus required identifying non-trivial entities, often spanning for many words, and on both winning solution is within 0.5–0.6 f-measure range. While not all business applications require entities of this type, such scenarios exist one shouldn’t expect CoNLL-2003 scores on such corpora.

In order to get a better understanding of the nature of our participants errors we decided to compute additional metrics: char-level F1-score (as opposed to span-level score reported previously). We observe that the most illustrative statistics is the difference between char-level F1-score and span-level F1-score. The average difference for top 3 participants is provided in Table 7 along with the mean length of each entity in chars.

Table 7: Differences in char-based f-measure and span-based

Metrics	ACT	BIN	CMP	ECO	INST	MET	QUA	SOC
Average F1 diff	0.28	0.03	0.00	0.23	0.21	0.27	0.00	0.19
Mean char length	34	12	10	24	27	31	12	21

One can easily notice that the difference is marginal for short entities and increases with the length of entity (moreover even the ordering of average F1 difference and mean char length is same with only one exception).

One possible explanation is that models have more difficulties with determining the exact borders of entities rather than detecting the entities themselves. With short entities there is little ground for border mistakes and the scores obtained are reasonably high. With longer entities the borders become less defined and thus the performance drops.

5.2. Methods

Since one of our main goals was to replicate a business scenario, we decided not to limit the participants in their choice of methods. They were at liberty to use any available methods, including proprietary models, as well as were allowed to create additional markup in order to train their model on a larger training set (participants were asked to send the organization committee any data they annotated themselves). All top participants, however, used exclusively open-source solutions and did not create any additional training data.

We have additionally published full unannotated corpus described in **Section 3.1** for the purpose of fine-tuning language models on it. To our best knowledge, however, no participant attempted it.

The methods used by most participants relied on academic standards.

For NER most participants started with popular CharCNN-BLSTM-CRF baseline [24] and attempted to improve it mainly with the help of contextualized word embeddings such as ELMo [31]. Two top systems are designed in essentially the same way: BERT [14] followed by MLP. The difference in scores between the two systems can be attributed to different BERTs used (multilingual BERT for the winner and RuBERT for the runner up) and different learning strategies.

Relation extraction allowed for better diversity of models. Several approaches were tested from simple heuristics and classical BLSTM-based approach [29] to once more BERT-derived pipelines. Unsurprisingly, the top two systems are both representatives of the latter category, however, unlike with NER the two systems have noticeable differences. The winner used R-BERT-inspired model [43]. Since R-BERT reimplementation is currently SOTA on SemEval-2010 Task 8, it is a small wonder, that its adaptation works well for Russian. The runner up has successfully reduced relation extraction to sequence-labelling task and employed multi-task learning simultaneously training on both NER and RE tasks.

6. Conclusion

In this paper we have presented RuREBus corpus and shared task.

Our main goal is to bridge the gap between academic corpora and real-world scenarios. Keeping it in mind, we have obtained a corpus of governmental texts, produced by the Ministry of Economic Development of the Russian Federation and developed a markup instruction for eight entity types and eleven relation types. We also provide a large (300 million tokens) corpus of unmarked texts of the same source, intended for language model training and fine-tuning.

Our corpus consists of texts with specific and non-trivial domain (i. e. governmental texts), containing non-perfect language and other irregularities. Our entities and relations are non-balanced and their spans can often be rather long. Thus in our opinion this corpus is well-suited for being test-case "worst-case" industrial application.

We have further organized a shared task on our corpus, thus establishing a reasonable baseline for it. The participating systems (8 for NER and 5 for RE) used methods, close to SOTA on academic baselines and yet were able to score rather unimpressive 0.56 for

NER task and 0.44 for RE. Given that simultaneously happening SemEval-2020 task 11 demonstrated comparable results, we can claim that this is the current performance on "worst-case" business scenarios. Indeed, often industrial application can contain both classical entities such as persons and entities similar to the ones present in our corpus, thus providing the scores in between traditional corpora and the ones recently developed.

Thus in our opinion we created a useful testing ground for applications of NER and RE in industry. We hope that it will be useful for NLP community in general and Russian NLP in particular.

Future work directions include but are not limited to developing more advanced machine learning methods and analytical solutions, better usage of linguistic feautures, ensembling of different approaches and open source tools.

7. Acknowledgements

Work on maintenance of the annotation system, discussions of results, and manuscript preparation was carried out by Elena Tutubalina, Vladimir Ivanov, Tatiana Batura and supported by the Russian Science Foundation grant no. 20-11-20166. Ekaterina Artemova and Veronika Sarkisyan worked on text annotation, discussions of results, and manuscript. Their work was supported by the framework of the HSE University Basic Research Program and Russian Academic Excellence Project “5-100”.

We would also like to thank students, participated in corpus annotation process, especially Anna Golenkovskaya, Polina Demina, Anastasia ..., Karolina ..., Ekaterina Sidorenko, Alexandra Pavlova.

References

1. *Alekseychuk, N. et al.:* Processing and analysis of russian strategic planning programs. In: International conference on digital transformation and global society. pp. 68–81 Springer (2019).
2. *Anh, L. T. et al.:* Application of a hybrid bi-lstm-crf model to the task of russian named entity recognition. Communications in Computer; Information Science book series—CCIS, volume 789 (2017).
3. *Anisimovich, K. et al.:* Syntactic and semantic parser based on abbyy compreno linguistic technologies. In: Computational linguistics and intellectual technologies: Proceedings of the international conference “dialog” [komp’iuternaia lingvistika i intellektual’nye tehnologii: Trudy mezhdunarodnoj konferentsii “dialog”]. pp. 90–103, Bekasovo, Russia (2012).
4. *Antonova, A. Y., Soloviev, A. N.:* Conditional random field models for the processing of russian. Communications of the ACM 56(6) (2013).
5. *Arkhipov, M. et al.:* Tuning multilingual transformers for named entity recognition on slavic languages. BSNLP’2019. 89 (2019).
6. *Artemova, E. et al.:* So what’s the plan? Mining strategic planning documents. In: Digital transformation and global society: Proceedings of the 5th international conference (dtgs 2020)., St. Petersburg, Russia (2020).

7. *AS, S. et al.:* FactRuEval 2016: Evaluation oF namEd entity recognition and fact extraction systems for russian.
8. *Bojanowski, P. et al.:* Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics. 5, 135–146 (2017).
9. *Brykina, M. M. et al.:* Dictionary-based ambiguity resolution in russian named entities recognition. International Workshop on Computational Linguistics; its Applications, ed. A. Narin'yani, v.1 (2013).
10. *Carreras, X., Màrquez, L.:* Introduction to the conll-2004 shared task: Semantic role labeling. In: In proceedings of conll2004. Association for Computational Linguistics, Boston, MA (2004).
11. *Chiticariu, L. et al.:* Rule-based information extraction is dead! Long live rule-based information extraction systems! In: Proceedings of the 2013 conference on empirical methods in natural language processing. pp. 827–832 (2013).
12. *Craidlin, L.:* Program of allocation of russian individualized nominal groups taglite. Computational linguistics; intellectual technologies Dialog (2005).
13. *Da San Martino, G. et al.:* SemEval-2020 task 11: Detection of propaganda techniques in news articles. In: Proceedings of the 14th international workshop on semantic evaluation., Barcelona, Spain (2020).
14. *Devlin, J. et al.:* BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019).
15. *Feldman, R., Rosenfeld, B.:* Boosting unsupervised relation extraction by using ner. In: Proceedings of the 2006 conference on empirical methods in natural language processing. pp. 473–481 (2006).
16. *Gareev, R. et al.:* Introducing baselines for russian named entity recognition. Computational Linguistics; Intelligent Text Processing (2013).
17. *Grishman, R., Sundheim, B.:* Design of the muc-6 evaluation. In: Proceedings of the 6th conference on message understanding. pp. 1–11 Association for Computational Linguistics, USA (1995).
18. *Hendrickx, I. et al.:* 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, Uppsala, Sweden (2010).
19. *Hovy, E. et al.:* OntoNotes: The 90. In: Proceedings of the human language technology conference of the naacl, companion volume: Short papers. pp. 57–60 Association for Computational Linguistics, USA (2006).
20. *Huang, Z. et al.:* Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991. (2015).
21. *Kutuzov, A., Kuzmenko, E.:* WebVectors: A toolkit for building web interfaces for vector semantic models. In: Ignatov, D. I. et al. (eds.) Analysis of images, social networks and texts: 5th international conference, aist 2016, yekaterinburg, russia, april 7–9, 2016, revised selected papers. pp. 155–161 Springer International Publishing, Cham (2017).

22. *Kuznetsov, A. et al.:* Family matters: Company relations extraction from wikipedia. In: International conference on knowledge engineering and the semantic web. pp. 81–92 Springer (2016).
23. *Lafferty, J. et al.:* Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning (2001).
24. *Lample, G. et al.:* Neural architectures for named entity recognition. In: Proceedings of the 2016 conference of the north American chapter of the association for computational linguistics: Human language technologies. pp. 260–270 Association for Computational Linguistics, San Diego, California (2016).
25. *Li, X. et al.:* A unified mrc framework for named entity recognition. ArXiv. abs/1910.11476, (2019).
26. *Ma, X., Hovy, E.:* End-to-end sequence labeling via bi-directional lstm-cnns-crf, (2016).
27. *Malykh, V., Lyalin, V.:* Named entity recognition in noisy domains. In: 2018 international conference on artificial intelligence applications and innovations (ic-aiai). pp. 60–65 IEEE.
28. *Malykh, V., Ozerin, A.:* Reproducing russian ner baseline quality without additional data. In: CDUD@ cla. pp. 54–59 (2016).
29. *Miwa, M., Bansal, M.:* End-to-end relation extraction using lstms on sequences and tree structures. 2016. arXiv preprint arXiv:1601.00770.
30. *Mozharova, V., Loukachevitch, N.:* Two-stage approach in russian named entity recognition. In: Intelligence, social media and web (ismw fruct), 2016 international fruct conference on. pp. 1–6 IEEE (2016).
31. *Peters, M. E. et al.:* Deep contextualized word representations. In: Proceedings of naacl-hlt. pp. 2227–2237 (2018).
32. *Piskorski, J. et al.:* 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 Association for Computational Linguistics, Florence, Italy (2019).
33. *Piskorski, J. et al.:* The first cross-lingual challenge on recognition, normalization, and matching of named entities in slavic languages. In: Proceedings of the 6th workshop on balto-slavic natural language processing. pp. 76–85 (2017).
34. *Podobryaev, A. V.:* Searching for person memories in news texts with the use of a model of conditional random fields. RCDL (2013).
35. *Popov, B. et al.:* KIM—a semantic platform for information extraction and retrieval. Journal of Natural Language Engineering 10 (2004).
36. *Shen, Y. et al.:* Deep active learning for named entity recognition. CoRR. abs/1707.05928, (2017).
37. *Soares, L. B. et al.:* Matching the blanks: Distributional similarity for relation learning. In: ACL. (2019).
38. *Stenetorp, P. et al.:* 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 Association for Computational Linguistics (2012).

39. *Tjong Kim Sang, E. F., De Meulder, F.:* 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. pp. 142–147 Association for Computational Linguistics, USA (2003).
40. *Trofimov, I.:* Identification of personal names in news texts on collections persons-1000/1111-f (in russian). Proceedings of RCDL-2014. 217–221 (2014).
41. *Tsygankova, T. et al.:* BSNLP2019 shared task submission: Multisource neural ner transfer. In: Proceedings of the 7th workshop on balto-slavic natural language processing. pp. 75–82 (2019).
42. *Vlasova, N. et al.:* Report on russian corpus for personal name retrieval. Proceedings of computational; cognitive linguistics TEL (2014).
43. *Wu, S., He, Y.:* Enriching pre-trained language model with entity information for relation classification. In: Proceedings of the 28th acm international conference on information and knowledge management. pp. 2361–2364 (2019).
44. *Zhang, Y. et al.:* 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 (2017).
45. *Zhao, Y. et al.:* Improving relation classification by entity pair graph. In: ACML. (2019).
46. *Zuev K. A., J. M. V., Indenbom M. E.:* StatiStical machine tranSlation with linguiStic language model. In: Computational linguistics and intellectual tech-nologies: Proceedings of the international conference “dialog” [komp’iuternaia lingvistika i intellektual’nye tehnologii: Trudy mezhdunarodnoj konferentsii “dialog”]. pp. 164–172, Bekasovo, Russia (2013).
47. *The ace 2005* (ace05) evaluation plan. (2005).