НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
июль–август 2023 Том 23 № 4 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
July–August 2023 Vol. 23 No 4 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2023-23-4-854-857

RuLegalNER: a new dataset for Russian legal named entities recognition

Zein Shaheen[1], Dmitry I. Mouromtsev[2]✉, Ignat Postny[3]

[1,2] ITMO University, Saint Petersburg, 197101, Russian Federation
[3] T.A.G. Consulting, Moscow, 119119, Russian Federation

[1] shaheen@itmo.ru, https://orcid.org/0000-0001-6802-2896
[2] mouromtsev@itmo.ru✉, https://orcid.org/0000-0002-0644-9242
[3] ipostny@gmail.com, https://orcid.org/0009-0005-9249-4160

Abstract

We address the scarcity of datasets specifically tailored for legal NER in the Russian language and investigate the generalization capabilities of models towards unseen named entities. A rule-based program developed by legal experts at Tag-Consulting Company was employed to automatically annotate legal texts and create the RuLegalNER dataset. Part of the named entities only exists in the development and test splits, and they are unseen in the training set. RuBERT was utilized as the base architecture for experimental evaluation. Two different architectural extensions were explored: RuBERT with CRF and RuBERT with adapters. These architectures were used to train and evaluate NER models on the RuLegalNER dataset. Utilize RuLegalNER to train and evaluate legal NER models, enhancing performance in the legal domain and studying generalization on unseen entities. A published version of RuLegalNER is presented with detailed statistics and demonstration of the usefulness of RuLegalNER by evaluating modern architectures.

Keywords

legal named entity recognition, natural language processing, information extraction, low-resource languages, transfer learning, transformers

For citation: Shaheen Z., Mouromtsev D.I., Postny I. RuLegalNER: a new dataset for Russian legal named entities recognition. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 4, pp. 854–857. doi: 10.17586/2226-1494-2023-23-4-854-857

УДК 004.912

RuLegalNER: новый датасет для распознавания именованных юридических сущностей на русском языке

Зейн Шахин[1], Дмитрий Ильич Муромцев[2]✉, Игнат Постный[3]

[1,2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[3] T.A.G. Consulting, Москва, 119119, Российская Федерация

[1] shaheen@itmo.ru, https://orcid.org/0000-0001-6802-2896
[2] mouromtsev@itmo.ru✉, https://orcid.org/0000-0002-0644-9242
[3] ipostny@gmail.com, https://orcid.org/0009-0005-9249-4160

Аннотация

Представлен новый датасет RuLegalNER, разработанный для обучения моделей распознавания именованных юридических сущностей на русском языке. Выполнена оценка способности моделей к обобщению при появлении в тексте ранее не встречавшихся именованных сущностей. Для автоматической разметки юридических текстов и создания набора данных RuLegalNER разработана программа на основе правил. Часть именованных сущностей в датасете была выделена в набор данных для валидации и тестирования и не встречается в обучающем наборе. Экспериментальная проверка датасета основана на базовой архитектуре RuBERT. Исследовано два расширения архитектуры: RuBERT с использованием CRF (Conditional Random Fields) и адаптеров. На основе архитектур выполнено обучение и оценка модели распознавания именованных сущностей на наборе данных RuLegalNER. Предложенный набор данных RuLegalNER может быть использован для создания новых моделей распознавания

© Shaheen Z., Mouromtsev D.I., Postny I., 2023

именованных сущностей в юридических текстах, что позволит автоматизировать контент-анализ юридических документов. Опубликована версия RuLegalNER с подробной статистикой и демонстрацией полезности набора данных RuLegalNER путем оценки на основе современных архитектур.

Ключевые слова
распознавание именованных юридических сущностей, обработка естественного языка, извлечение информации, языки с ограниченными ресурсами, передаточное обучение, трансформеры

Ссылка для цитирования: Шахин З., Муромцев Д.И., Постный И. RuLegalNER: новый датасет для распознавания именованных юридических сущностей на русском языке // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23, № 4. С. 854–857 (на англ. языке). doi: 10.17586/2226-1494-2023-23-4-854-857

Recognizing named entities in legal texts is a crucial task in natural language processing, with applications ranging from information extraction [1] to legal research [2]. However, the availability of resources specifically tailored for legal Named Entity Recognition (NER) in the Russian language is limited. This scarcity poses a significant challenge for researchers and practitioners working in the legal domain. Additionally, annotating legal datasets with expert human annotators is an expensive process. Furthermore, even with expert annotations, noisy labels can be present due to the complexity and ambiguity of legal texts [3].

In the Russian language, the availability of datasets specifically focused on legal named entity recognition is scarce. Among the limited options [4–6], one notable dataset is NEREL [7] which includes a total of 20 classes, with only three classes specifically related to the legal domain: LAW, CRIME, and PENALTY. However, out of the 56,000 annotated named entity instances in NEREL, only 1,679 pertain to legal named entities. This highlights the need for more comprehensive and domain-specific resources in the legal domain for Russian.

To address these challenges, this paper introduces RuLegalNER, a rule-based annotated legal dataset for the Russian language. The dataset was created using a rule-based program developed by legal experts in Tag-Consulting Company, enabling the automatic annotation of legal named entities in a large collection of Russian legal texts. This rule-based approach alleviates the need for extensive manual annotations by experts, speeding up the dataset creation process. Although rule-based annotation may introduce some noise, it serves as a starting point for subsequent refinement and iterative improvement.

To study the generalization ability of a named entity recognition model trained on an automatically annotated dataset, we developed RuLegalNER, a dataset of Russian legal documents. This dataset is annotated with more than 20 classes of named entities. However, it is important to note that not all legal named entities present in the documents are annotated, and the annotation coverage is sparse. From the initial set of classes, we specifically selected five classes for inclusion in this dataset: Individual person, legal entity, Penalty, Crime, and Law. The annotation process was performed using a rule-based system provided by TAG Consulting Company.

To ensure the evaluation of the model performance on unseen named entities, we incorporated low frequency entities into the dataset. These entities were treated as unseen during the training process, and they were exclusively reserved for the validation and test stages.

The RuLegalNER dataset consists of a sample of 100,000 Russian legal documents. Within this dataset, there are a total of 860 unique named entities. Notably, 289 of these entities appear only in the test set, resulting in a total of 777 occurrences of unseen entities in the test set.

For detailed statistics on the distribution of named entities in each split of the dataset, please refer to Table 1. The table shows the number of unique entities and their frequencies within each portion of the dataset as well as statistics for both seen and unseen entities in the test set. Additionally, Figure provides samples from the dataset, showcasing the variety of named entities present. The dataset is publicly available[1].

We evaluated our models ability to predict unseen named entities, handle misspellings, and different grammatical cases. Objective metrics, such as precision, recall, and f1 score, were employed to assess their prediction power. We employed additional two objective evaluation metrics: Count of Predicted Unseen Named Entities (CP-UNE) and Count of Unique Predicted Unseen Named Entities (CUP-UNE), to measure the models generalization ability to unseen named entities.

In our experiments, we utilized various NER models for our research. The first model, RuBERT-NER, is based on RuBERT [8], a Russian text feature extraction model trained using the Russian version of Wikipedia and multilingual-BERT as the base checkpoint. We fine-tuned RuBERT using legal documents and employed a token classifier to generate probabilities for different classes. An extension of RuBERT-NER, RuBERT-NER-CRF, incorporated Conditional Random Fields (CRFs) to capture long-range dependencies and improve prediction accuracy. It utilized the Viterbi algorithm and a learnable state-state transition matrix for decoding the output labels. Another extension, RuBERT-NER-Adapter, employed adapters [9], a transfer learning strategy, to augment RuBERT-NER without significantly increasing the number of parameters. Lastly, we used a baseline model, BiLSTM-CRF [10], which combined a bidirectional LSTM with a CRF component for NER tasks. This model was pretrained on Russian Wikipedia and fine-tuned using legal documents.

Given the sparse nature of the annotated dataset and to evaluate the performance of these models and make comparisons, we divided the legal documents into segments, each containing 60 words, and filtered out segments that did not contain any annotated legal entities. The remaining segments were utilized for training,

[1] Available at: https://github.com/zeino8/RuLegalNER (accessed: 18.07.2023).

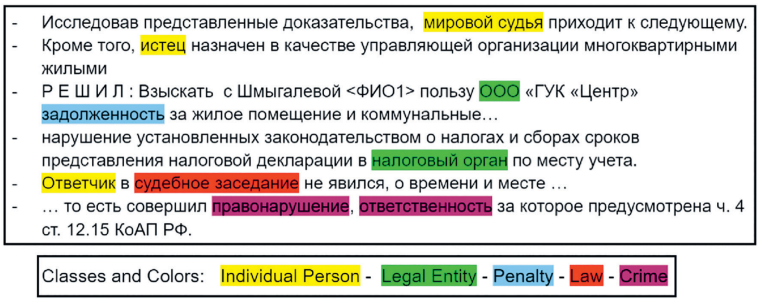

Figure. Legal texts samples from RuLegalNER with classes and their colors, and entities highlighted with the corresponding class color

Table 1. Named entity data in the RuLegalNER dataset organized by class number

Class	# unique entities					# occurrences				
	Train	Validation	Test		Dataset	Train	Validation	Test		Dataset
			Seen	Unseen				Seen	Unseen	
Individual	320	521	293	192	640	214,591	44,553	32,893	258	291,295
Legal Entity	30	55	30	26	59	28,343	6922	4599	165	40,029
Penalty	35	66	35	31	70	70,224	15,666	10,981	169	97,040
Crime	21	38	21	17	42	61,158	12,766	8897	91	82,912
Law	25	45	25	23	49	91,771	17,727	12,483	94	122,075
Total	431	725	404	289	860	466,087	97,634	68,853	777	633,351

Table 2. Comparing objective evaluation metrics (precision, recall, f1) for various models: (1) RuBERT-NER, (2) RuBERT-NER-CRF, (3) RuBERT-NER-Adapter, and (4) BiLSTM-CRF on all classes together. We also count number of previously unseen named entities (CP-UNE) and number of unique previously unseen named entities (CUP-UNE)

Model	Precision	Recall	F1-score	CP-UNE	CUP-UNE
RuBERT-NER	0.951	0.960	0.956	58	46
RuBERT-NER-CRF	0.976	0.847	0.907	57	48
RuBERT-NER-Adapter	0.905	0.940	0.922	52	40
BiLSTM-CRF	0.937	0.939	0.938	7	3

validation, and testing purposes. Evaluation results and comparison between models is presented in Table 2.

We introduced RuLegalNER dataset, a legal named entity recognition dataset in the Russian language. With its automatically annotated legal documents, RuLegalNER provides a valuable resource for training and evaluating NER systems in the legal domain. The evaluation of various modern architectures for NER models on RuLegalNER highlights the strengths and limitations of each approach, enabling researchers to make informed decisions regarding model selection.

References

1. Weston L., Tshitoyan V., Dagdelen J., Kononova O., Trewartha A., Persson K.A., Ceder G., Jain A.. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *Journal of Chemical Information and Modeling*, 2019, vol. 59, no. 9, pp. 3692–3702. https://doi.org/10.1021/acs.jcim.9b00470
2. Angelidis I., Chalkidis I., Koubarakis M. Named entity recognition, linking and generation for greek legislation. *Legal Knowledge and Information Systems*, 2018, vol. 313, pp. 1–10.
3. Zhu Y., Ye Y., Li M., Zhang J., Wu O. Investigating annotation noise for named entity recognition. *Neural Computing and Applications*, 2023, vol. 35, no. 1, pp. 993–1007. https://doi.org/10.1007/s00521-022-07733-0
4. Vlasova N.A., Suleymanova E.A., Trofimov I.V. Report on Russian corpus for personal name retrieval. *Proceedings of Computational and Cognitive Linguistics, TEL*, 2014, pp. 36–40.

Литература

1. Weston L., Tshitoyan V., Dagdelen J., Kononova O., Trewartha A., Persson K.A., Ceder G., Jain A.. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature // Journal of Chemical Information and Modeling. 2019. V. 59. N 9. P. 3692–3702. https://doi.org/10.1021/acs.jcim.9b00470
2. Angelidis I., Chalkidis I., Koubarakis M. Named entity recognition, linking and generation for greek legislation // Legal Knowledge and Information Systems. 2018. V. 313. P. 1–10.
3. Zhu Y., Ye Y., Li M., Zhang J., Wu O. Investigating annotation noise for named entity recognition // Neural Computing and Applications. 2023. V. 35. N 1. P. 993–1007. https://doi.org/10.1007/s00521-022-07733-0
4. Vlasova N.A., Suleymanova E.A., Trofimov I.V. Report on Russian corpus for personal name retrieval // Proceedings of Computational and Cognitive Linguistics, TEL. 2014. P. 36–40.

5. Starostin A.S., Bocharov V.V., Alexeeva S.V., Bodrova A.A., Chuchunkov A.S., Dzhumaev S.S., Efimenko I.V., Granovsky D.V., Khoroshevsky V.F., Krylova I.V., Nikolaeva M.A., Smurov I.M., Toldova S.Y. Factrueval 2016: evaluation of named entity recognition and fact extraction systems for Russian. *Proc. of the International Conference "Dialogue 2016"*, 2016, pp. 702–720.
6. Gareev R., Tkachenko M., Solovyev V., Simanovsky A., Ivanov V. Introducing baselines for russian named entity recognition. *Lecture Notes in Computer Science*, 2013, vol. 7816, pp. 329–342. https://doi.org/10.1007/978-3-642-37247-6_27
7. Loukachevitch N., Artemova E., Batura T., Braslavski P., Denisov I., Ivanov V., Manandhar S., Pugachev A., Tutubalina E. Nerel: A Russian dataset with nested named entities, relations and events. *Proc. of the Recent Advances in Natural Language Processing*, 2021, pp. 876–885 https://doi.org/10.26615/978-954-452-072-4_100
8. Kuratov Y., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019"*, 2019.
9. Houlsby N., Giurgiu A., Jastrzebski S., Morrone B., De Laroussilhe Q., Gesmundo A., Attariyan M., Gelly S. Parameter-efficient transfer learning for NLP. *Proc. of the 36th International Conference on Machine Learning*, 2019, pp. 2790–2799.
10. Panchendrarajan R., Amaresan A. Bidirectional LSTM-CRF for named entity recognition. *Proc. of the 32nd Pacific Asia Conference on Language, Information and Computation*, 2018, pp. 531–540.

5. Starostin A.S., Bocharov V.V., Alexeeva S.V., Bodrova A.A., Chuchunkov A.S., Dzhumaev S.S., Efimenko I.V., Granovsky D.V., Khoroshevsky V.F., Krylova I.V., Nikolaeva M.A., Smurov I.M., Toldova S.Y. Factrueval 2016: evaluation of named entity recognition and fact extraction systems for Russian // Proc. of the International Conference "Dialogue 2016". 2016. P. 702–720.
6. Gareev R., Tkachenko M., Solovyev V., Simanovsky A., Ivanov V. Introducing baselines for russian named entity recognition // Lecture Notes in Computer Science. 2013. V. 7816. P. 329–342. https://doi.org/10.1007/978-3-642-37247-6_27
7. Loukachevitch N., Artemova E., Batura T., Braslavski P., Denisov I., Ivanov V., Manandhar S., Pugachev A., Tutubalina E. Nerel: A Russian dataset with nested named entities, relations and events // Proc. of Recent Advances in Natural Language Processing. 2021. P. 876–885 https://doi.org/10.26615/978-954-452-072-4_100
8. Kuratov Y., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019". 2019.
9. Houlsby N., Giurgiu A., Jastrzebski S., Morrone B., De Laroussilhe Q., Gesmundo A., Attariyan M., Gelly S. Parameter-efficient transfer learning for NLP // Proc. of the 36th International Conference on Machine Learning. 2019. P. 2790–2799.
10. Panchendrarajan R., Amaresan A. Bidirectional LSTM-CRF for named entity recognition // Proc. of the 32nd Pacific Asia Conference on Language, Information and Computation. 2018. P. 531–540.

Authors

Zein Shaheen — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57209279132, https://orcid.org/0000-0001-6802-2896, shaheen@itmo.ru
Dmitry I. Mouromtsev — PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 55575780100, https://orcid.org/0000-0002-0644-9242, mouromtsev@itmo.ru

Ignat Postny — Director, T.A.G. Consulting, Moscow, 119119, Russian Federation, https://orcid.org/0009-0005-9249-4160, ipostny@gmail.com

Авторы

Шахин Зейн — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57209279132, https://orcid.org/0000-0001-6802-2896, shaheen@itmo.ru
Муромцев Дмитрий Ильич — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 55575780100, https://orcid.org/0000-0002-0644-9242, mouromtsev@itmo.ru
Постный Игнат — директор, T.A.G. Consulting, Москва, 119119, Российская Федерация, https://orcid.org/0009-0005-9249-4160, ipostny@gmail.com

Received 19.06.2023
Approved after reviewing 23.06.2023
Accepted 30.07.2023

Статья поступила в редакцию 19.06.2023
Одобрена после рецензирования 23.06.2023
Принята к печати 30.07.2023

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»