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

Moscow, June 17–20, 2020

RELATION EXTRACTION DATASET FOR THE RUSSIAN

Gordeev D. I. (gordeev-di@ranepa.ru),
Davletov A. A. (davletov-aa@ranepa.ru),
Rey A. I. (rey-ai@ranepa.ru),
Akzhigitova G. R. (akzhigitova-gr@ranepa.ru),
Geymbukh G. A. (geymbukh-ga@ranepa.ru)
RANEPA, Moscow, Russia

There are few existing relation extraction datasets for the Russian language and they contain a rather small number of examples. Thus, we decided to create a new Ontonotes-based named entities and relation extraction sentence-level dataset called RURED. The dataset contains more than 500 annotated texts and more than 5,000 labelled relations. We also publish baseline models for relation extraction and named entity recognition trained on the dataset. Our models achieve 0.85 for named entity recognition and 0.78 for relation extraction in F1-score.[1]

Key words: relation extraction, named entity recognition, tacred, bert

DOI: 10.28995/2075-7182-2020-19-348-360

ДАТАСЕТ ДЛЯ ИЗВЛЕЧЕНИЯ ОТНОШЕНИЙ ДЛЯ РУССКОГО ЯЗЫКА

Гордеев Д. И. (gordeev-di@ranepa.ru),
Давлетов А. А. (davletov-aa@ranepa.ru),
Рей А. И. (rey-ai@ranepa.ru),
Акжигитова Г. Р. (akzhigitova-gr@ranepa.ru),
Геймбух Г. А. (geymbukh-ga@ranepa.ru)
РАНХиГС, Москва, Россия

[1] https://github.com/InstituteForIndustrialEconomics/rured

На данный момент существует немного размеченных наборов данных для извлечения отношений из текстов. В данный статье мы представляем такой датасет RURED, содержащий разметку именованных сущностей по схеме Ontonotes и отношений между ними на уровне предложений. Датасет содержит более 500 аннотированных текстов и более 5000 размеченных отношений. Также мы публикуем основанные на BERT модели, обученные на этом наборе данных. В задаче автоматического распознавания именованных сущностей модель достигла 0,85 п. п. по метрике F1, для задачи извлечения отношений — 0,78.

Ключевые слова: извлечение отношений, распознавание именованных сущностей, tacred, bert

1. Introduction

The task of relation extraction is to find entities in a sentence and establish the type of relations between them, i.e. to extract triplets from texts: (entity 1; entity 2; their relationship). For example, in the sentence “Mark Zuckerberg, the founder of Facebook, bought a startup.” there are named entities: “Mark Zuckerberg” and “Facebook”, which are connected by the relation “Founder”. Relation extraction is useful for building taxonomies and extracting facts from texts.

There are several approaches to the problem:

- supervised learning-based methods
- distant supervision-based methods

A popular approach to relation extraction is distant supervision [13]. This method uses an ontology database and a large text corpus to align sentences containing entities. Unfortunately, this method is prone to noisy labels [12]. There have been numerous attempts at fixing problems of distant supervision. However, still, most models ignore categories in the long tail of the distribution [6]. Moreover, classes tend to be of some distinct domains that are typical of the database (e.g. these are locations and nationalities in the case of NYT10 [16] built on top of Freebase [1]). If we are interested in relations that are absent from the database, we have no choice but to resort to supervised or semi-supervised methods.

There is a decent number of named entity datasets for the Russian language. There are traditional Person-Organization-Location (e.g. [5], [7], [14]) datasets as well as more specialized ones that are devoted to a single type of entities (e.g. Persons-1000[2]). Moreover, it is possible to use transfer learning and zero-shot learning for named entity recognition. It was shown that a multilingual BERT-based model which is fine-tuned on an English NER (named entity recognition) dataset is able to gain reasonable results for Russian[3] (unrelated to this work, but we tried a similar transfer learning approach

[2] http://ai-center.botik.ru/Airec/index.php/ru/collections/28-persons-1000

[3] http://docs.deeppavlov.ai/en/master/features/models/ner.html#multilingual-bert-zero-shot-transfer

with TACRED and it did not bring us any results. We will explore it in future work). However, for relation extraction the situation is more challenging.

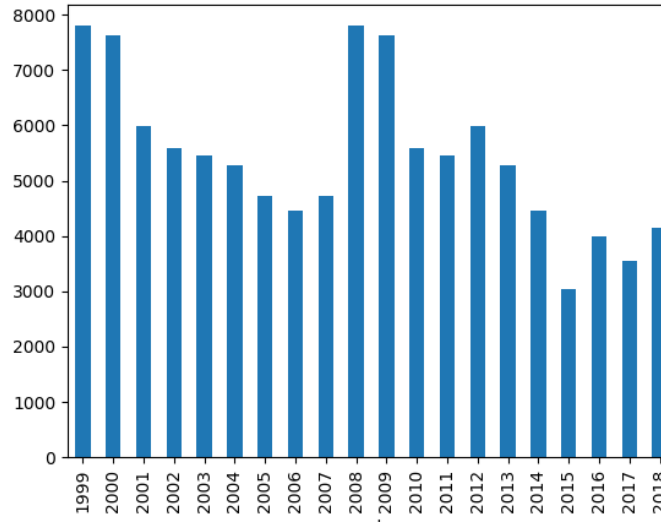

Figure 1: Yearly distribution of economy articles in Lenta.ru

Supervised approaches often treat the problem of relation extraction as a classification task. This approach allows for achieving higher quality predictions. Moreover, the classes may be beyond the scope of knowledge bases. However, the annotation procedure is demanding and tiresome. That is why such datasets are few and exist only for major languages. For the Russian language the only existing annotated dataset was published for the competition FactRuEval 2016 [17] held in conjunction with the conference Dialogue 2016. It contains 1059 facts (which may contain multiple relations). It is often not enough for training a classifier. Even the organizers of FactRuEval 2016 state that "the small size of the demo corpus shut out systems that relied on machine learning and made it difficult to fine-tune rule-based systems". The best performing system achieved the F1 score of 0.51. At the same time as our work there was created RuREBus dataset[4] which was used for RuREBus shared task held together with Dialog 2020 conference. It contains about 300 annotated texts from the domain of Russian municipal legal documents. It can be used as a benchmark for relation extraction and named entity recognition algorithms and it is close in its nature to business domains. However, due to the specificity of the domain it cannot be used for general purposes. Another interesting article focused on extracting relations from Wikipedia pages [10]. It contained both an automatically extracted dataset and manually labelled data. The annotators linked only relations between the 'main' company (the topic of the Wikipedia page) and already highlighted mentions of other companies (which have a corresponding Wikipedia page or it is to be created). This procedure works only for Wikipedia and similar

[4] https://github.com/dialogue-evaluation/RuREBus/

domains because we do not usually know the topic of the text and there may be relations between non-topical entities. Thus, we decided to create RURED (RUssian Relation Extraction Dataset). It contains 536 annotated texts and 5,381 relations. The number of labelled named entities is 22,595. Using training data from the dataset we trained several models for named entity recognition and relation extraction.

2. Dataset

We used Lenta.ru news corpus for annotation.[5] Only texts with tag "Экономика" (economy) were selected because we were mostly interested in economic events for our future work. Lenta.ru news dataset contains news articles from 1999 till 2019. All texts were selected randomly for annotation.

2.1. Named entities labelling

Named entities were automatically annotated using BERT [3] model trained on English Ontonotes [4] provided by DeepPavlov [2]. The cross-lingual nature of BERT allows us to successfully infer named entities for the Russian language despite the model being fine-tuned on the English dataset. During annotation, named entities were manually corrected if wrong and new entities were added when necessary.

We adhered to Ontonotes 5.0 guidelines and stuck to its annotation procedure.[6] Several new types and subtypes of named entities were also added (see **Table 1**). For example, GPE was split into several separate types: COUNTRY, CITY, REGION, BOROUGH.

Table 1: New named entity types besides those in Ontonotes

Named entity type	Subtype-of	Description
PROFESSION	—	Professions and people of these professions. Corresponds to 'title' in TACRED
COUNTRY	GPE	Names of countries
REGION	GPE	Names of sub-country entities
CITY	GPE	Names of cities, towns and villages
BOROUGH	GPE	Names of sub-city entities
GROUP	—	unnamed groups of people and companies
FAMILY	GROUP	families as a whole
AGE	NUMBER (not used in annotation)	people's and objects' ages
NATIONALITY	NORP	names of nationalities
RELIGION	NORP	names of religions
CURRENCY	—	names of currencies

[5] https://github.com/yutkin/Lenta.Ru-News-Dataset

[6] https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf

Nested named entities were not labelled—only upper-level entities were annotated. For example, the whole phrase ‘посол РФ в Камеруне’ (ambassador of Russia to Cameroon) was labelled as ‘profession’, child entities ‘РФ’ (Russian Federation) and ’Камеруне’ (Cameroon) were not labelled. It might be changed in future releases.

Prepositions were included in named entities (usually dates and numbers) if they were vital for the entity (e.g. “since 1992” was labelled as a single entity). However, it resulted in some confusion among annotators due to verb and nouns valency.

State and organization departments were labelled as “ORGANIZATIONS” and later were connected with “OWNERSHIP” relation (Fig. 4).

Unfortunately, as with many natural datasets we see a heavy-tailed class distribution here (Fig. 2).

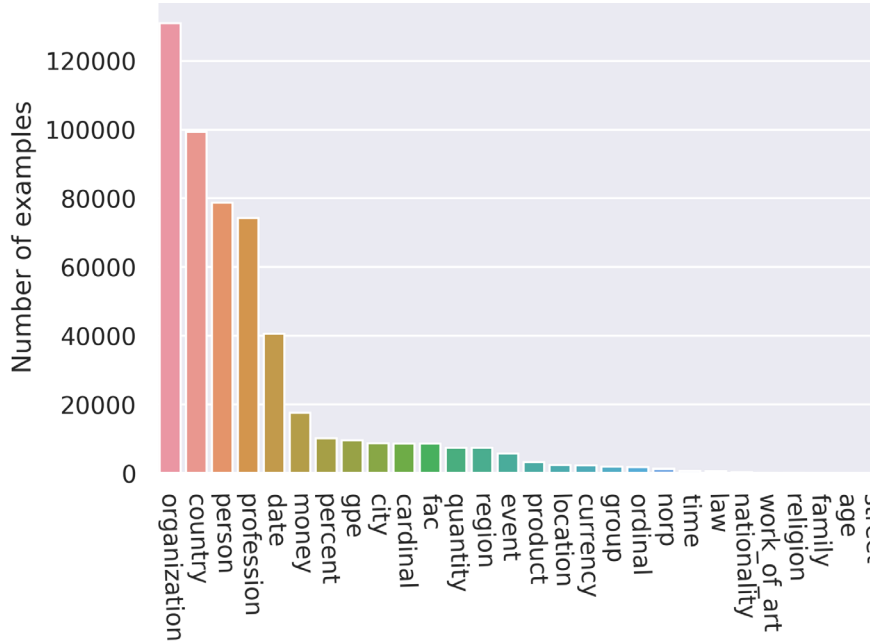

Figure 2: Distribution of named entity classes

2.2. Relation extraction labelling

Our relation extraction dataset is based on TACRED (The TAC Relation Extraction Dataset) [20]. TACRED is an English-language data corpus with labelled entities and relations between them. The markup was done manually using the MTurk HIT distributed annotation system based on data from TAC KBP 2009–2015 competitions. In the competition, each participant is given 100 entities and a large body of texts containing them. Participant systems must extract heterogeneous attributes for each entity and use them to fill a knowledge base. The knowledge base consists

of named entities and their attributes. TACRED contains 42 types of relations. 79.9% of examples are of the class 'no relation'. There are 106,264 examples in the dataset. One of the downsides of TACRED is that approximately 90% of sentences contain only a single annotated relation despite there being multiple possible relations. For unknown texts, this approach also requires to check all named entities against all other named entities ($O(N^2)$ complexity) because relations are intransitive.

To overcome this problem we labelled all relations in a sentence. It allows training models that predict all possible relations for the word in a single run ($O(N)$).

Brat annotation tool [18] has been used for relation and named entities annotation.

All relations were labelled if the fact has been true at some point in time (e.g. if the person is no longer working for the company, but worked for it in the past, the relation is considered valid). It correlates with annotation procedures by other researchers [6].

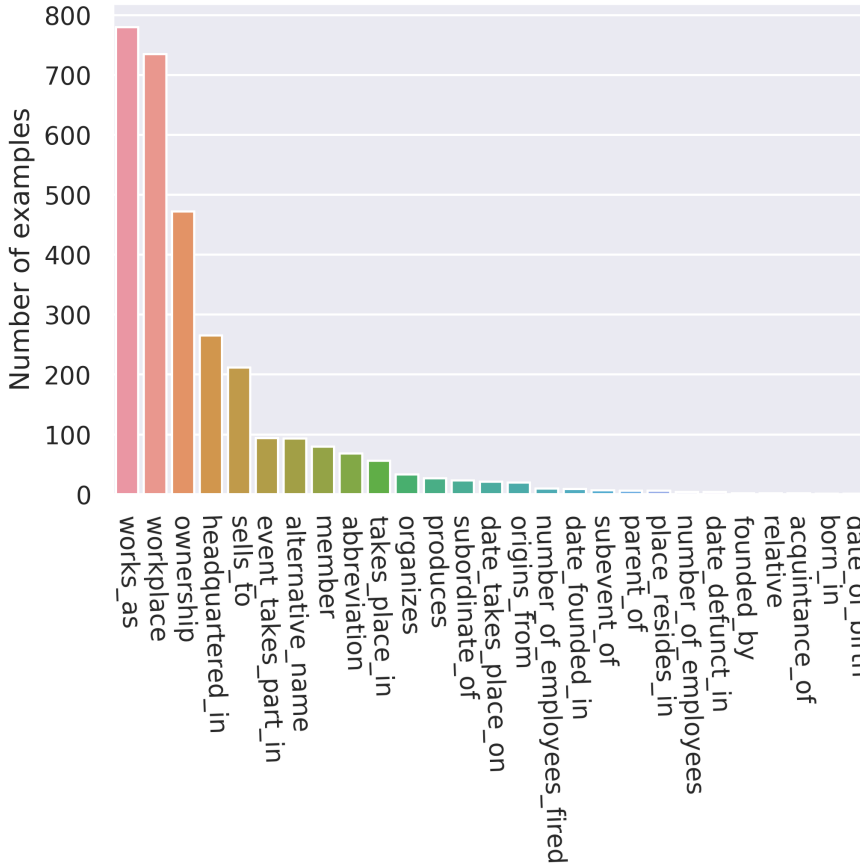

Figure 3: Distribution of relation classes

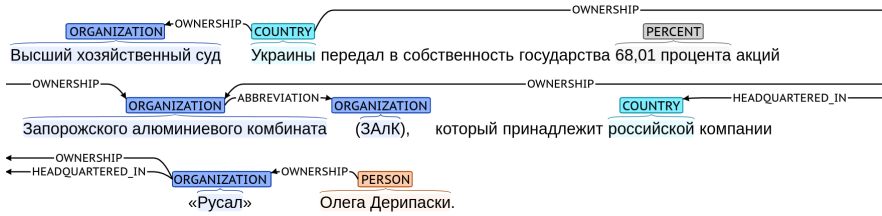

Figure 4: Annotation example. Child organizations and governmental departments are labelled as distinct entities and are connected with relation "OWNERSHIP". Relations are labelled with disregard to time. Both Ukraine and Rusal own ZAlK

The labelling was performed at the level of sentences. We did not want to complicate the annotation by reference resolution. Moreover, there are existing reference resolution datasets [11] for Russian which can be used together with this corpus. However, in some cases, annotators labelled relations across sentences. They were not removed as they can be easily deleted by post-processing.

Table 2: Relation types

Relation	Parent NERs	Child NERs	Subtype of
PRODUCES	FACILITY, ORGANIZATION	PRODUCT	—
TAKES PLACE IN	EVENT	FAC, ORGANIZATION, GPE	—
DATE TAKES PLACE ON	EVENT	DATE	—
ORGANIZES	GROUP, ORGANIZATION, GPE	EVENT	—
EVENT TAKES PART IN	GROUP, ORGANIZATION, PERSON, GPE	EVENT	—
NUMBER OF EMPLOYEES	FAC, ORGANIZATION	CARDINAL, QUANTITY	—
NUMBER OF EMPLOYEES HIRED	FAC, ORGANIZATION	CARDINAL, QUANTITY	—
NUMBER OF EMPLOYEES FIRED	FAC, ORGANIZATION	CARDINAL, QUANTITY	—
HEADQUARTERED IN	FAC, ORGANIZATION	LOCATION, GPE	—
WORKS AS	PERSON, GROUP	PROFESSION	—
WORKPLACE	PROFESSION, PERSON, GROUP	ORGANIZATION, GPE	—
SUBORDINATE OF	PROFESSION, PERSON, GROUP	PROFESSION, PERSON	—

Relation	Parent NERs	Child NERs	Subtype of
ACQUAINTANCE OF	PERSON, GROUP	PERSON	—
FRIEND OF	PERSON, GROUP	PERSON	ACQUAINTANCE OF
RELATIVE	PERSON, GROUP	PERSON	ACQUAINTANCE OF
PARENT OF	PERSON, GROUP	PERSON	ACQUAINTANCE OF
SIBLING	PERSON	PERSON	ACQUAINTANCE OF
MEMBER	COUNTRY, REGION	GPE	—
OWNERSHIP	LOCATION	LOCATION	—
SELLS TO	GROUP, ORGANIZATION, PERSON, GPE	GROUP, ORGANIZATION, PERSON, GPE	—
ALTERNATIVE NAME	ORGANIZATION	GPE	—
ABBREVIATION	FAC	FAC	ALTERNATIVE NAME
FOUNDED BY	ORGANIZATION	PERSON, ORGANIZATION, FAMILY, GROUP	—
ORIGINS FROM	PERSON, FAMILY, GROUP	NATIONALITY, GPE	—
PLACE RESIDES IN	PERSON	LOCATION, GPE	—
DATE FOUNDED IN	ORGANIZATION, GPE	DATE	—
DATE DEFUNCT IN	ORGANIZATION, GPE	DATE	—
DATE OF DEATH	PERSON	DATE	—
DATE OF BIRTH	PERSON	DATE	—
AGE IS	PERSON	AGE	—
AGE DIED AT	PERSON	AGE	—
BORN IN	PERSON	GPE	—
PLACE OF DEATH	PERSON	GPE, LOCATION	—
SUBEVENT OF	EVENT	EVENT	—

The relation annotation was first individually performed by one of two economists. Afterwards, each annotated text was reviewed and fixed if necessary by one of two linguists to improve annotation quality. Difficult cases were discussed together.

All in all 536 texts were annotated. They contain 6,931 sentences in total, 2,330 of which contain a relation. The average text length is 288 words. In total these texts contain 5,381 relations and 22,595 distinct named entities. The dataset contains 22,846 unique tokens. Class distribution can be seen in the picture (**Fig. 3**).

We followed the MAMA cycle for dataset labelling [15]. During the annotation procedure, we weekly trained SpanBERT models [8] and inferred predictions for classes with good precision (higher than 0.9) on not-yet annotated data. During inference, only relations with softmax scores higher than 0.9 were selected. This procedure helped us to accelerate annotation, yet it brought some mistakes.

2.3. Inter-annotator agreement

We used Cohen's Kappa for inter-annotator agreement measurement. For named entities its value is 0.77; for relations it is 0.79. Most disagreements were connected with named entity spans. In many cases, relations labelling disagreement can also be attributed to named entities mismatches (see Fig. 5). It was sometimes the case that the first annotator mistakingly labelled named entity as a whole while our guidelines require to split it into several entities and label a relation between them (if there exists one). Also introducing hierarchy in relations and named entities lead to some disagreement between annotators.

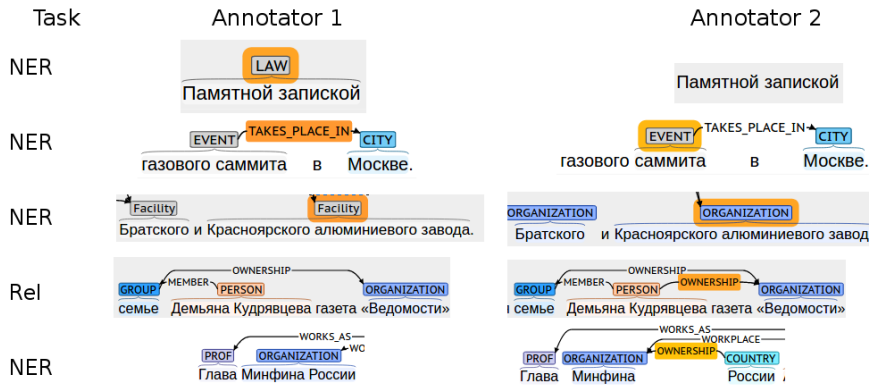

Figure 5: Some disagreement examples

2.4. Annotation Challenges

In many cases, it was difficult to label relations since we did not perform coreference resolution and examined only single sentences. If we had labelled references (e.g. pronouns) as named entities, it would have hindered named entities recognition quality. If we had labelled relations across sentences, it would require to choose among many mentions of the corresponding named entity from the text (sometimes it is not obvious). Moreover, in this case relations tend to span across many sentences which might impose a major difficulty for modern systems [19]. Thus, we decided to skip relations where both entities could not be labelled within a single sentence. In future, we are going to change it.

3. Training models

Using the annotated dataset we trained models for named entity recognition and relation extraction. For all models we randomly separated the sentences from the dataset into train, validation and test datasets using the 0.8: 0.1: 0.1 ratio with disregard to their date and context. We also attempted at fine-tuning them with various learning rates.

3.1. Named entity recognition

For named entity recognition we used a model based on the multilingual BERT model. The BERT [3] system is a transformer-based [19] model that has been pre-trained on a huge text dataset. In the course of pre-training, the tasks of the next sentence prediction and masked language modelling were solved jointly. During the next sentence prediction task, given two input sentences the system determined whether the second sentence is a continuation of the first. In the language modelling problem, the task was to predict masked words using all other words from the sentence. 15% of the words in the original sentence are replaced with a special token [MASK] and, the system predicts words at the positions of the masks. For the next sentence prediction task, a vector representation corresponding to the sentence start token [CLS] was used. h For named entity recognition, we use last layer hidden state of the BERT encoder. Afterwards it is passed through a softmax layer. We use cross entropy as our loss function.

To distinguish between the considered entities in the sentence, they are isolated with special tokens representing the beginning and the end of the entity.

3.2. Relation extraction

In this paper, we treat relation extraction as a classification problem. There is a large body of training examples. Each training example is a sentence with a pair of entities and the relationship between them. “No relation” is one of the relations classes.

In this work we use code provided by SpanBERT [7] which is a relation extraction system based on BERT. SpanBERT demonstrated near-SOTA (state-of-the-art) results on the TACRED [21] dataset. Given a sentence, two entities from it and the relation between them (‘no relation’ is a possible outcome) named entities are replaced with their NER-tags. A linear classifier is added on top of [CLS] token to predict the relation type. We also tried replacing BERT-weights with a pretrained RuBERT-model which is finetuned on Russian texts [9]. However, it did not improve our results.

Table 3: Relation extraction and named entity recognition results

task/model	learning rate	dataset	f1	precision	recall
named entities recognition	**1e-05**	**test**	**0.85**	**0.83**	**0.86**
relation extraction	1e-05	dev	0.807	0.861	0.760
relation extraction	2e-05	dev	0.805	0.843	0.769
relation extraction	**1e-05**	**test**	**0.782**	**0.841**	**0.731**
relation extraction	2e-05	test	0.778	0.814	0.744
relation extraction / rubert	1e-05	dev	0.813	0.861	0.77
relation extraction / rubert	1e-05	test	0.779	0.825	0.739
relation extraction / rubert	1e-05	dev	0.813	0.861	0.77
relation extraction / rubert	1e-05	test	0.779	0.825	0.739

Results for relation extraction and named entity recognition are provided in Table 3.

4. Conclusions

In this paper, we have introduced our new dataset containing named entities and relations between them. We also published baseline models that can be used by practitioners and researchers. We hope that this work will enhance NLP-research for the Russian language and will serve as a public baseline for future research on NER and relation extraction. We also hope that it will be of use to the non-scientific NLP community. In future work we would like to broaden dataset domain and to make it even more representative.

5. Licensing

Annotations and trained models are published under MIT license. Lenta.ru news articles are the property of the corresponding copyright holders.

References

1. *Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor.* Freebase: A collaboratively created graph database for structuring human knowledge. In Proceedings of the ACM SIGMOD International Conference on Management of Data, pages 1247–1249, 2008.
2. *Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva, and Marat Zaynutdinov.* DeepPavlov: Open-Source library for dialogue systems. In ACL 2018—56th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations, pages 122–127, 2018.
3. *Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.* BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. oct 2018.
4. *Ralph Weischedel et Al.* OntoNotes Release 5.0 LDC2013T19. Linguistic Data Consortium, 2013.
5. *Rinat Gareev, Maksim Tkachenko, Valery Solovyev, Andrey Simanovsky, and Vladimir Ivanov.* Introducing baselines for russian named entity recognition. In International Conference on Intelligent Text Processing and Computational Linguistics , pages 329–342. Springer, 2013.
6. *Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy.* SpanBERT: Improving Pre-training by Representing and Predicting Spans. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 13881398, Florence, Italy, 2019. Association for Computational Linguistics.

7. *Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert:* Improving pre-training by representing and predicting spans. arXiv preprint arXiv:1907.10529, 2019.
8. *Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, Omer Levy, and y Allen.* SpanBERT: Improving Pre-training by Representing and Predicting Spans. Technical report.
9. *Yu. Kuratov and M. Arkhipov.* Adaptation of deep bidirectional multilingual transformers for Russian language. In Komp'juternaja Lingvistika i Intellektual'nye Tehnol., volume 2019-May, pages 333–339, 2019.
10. *Artem Kuznetsov, Pavel Braslavski, and Vladimir Ivanov. Family matters:* Company relations extraction from wikipedia. In International Conference on Knowledge Engineering and the Semantic Web, pages 81–92. Springer, 2016.
11. *T. A. Le, M. A. Petrov, Y. M. Kurato, and M. S. Burtsev.* Sentence Level Representation and Language Models in The Task of Coreference Resolution for Russian. In Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2019), pages 341–350, 2019.
12. *Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek.* Distant Supervision for Relation Extraction with an Incomplete Knowledge Base. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 777–782, Atlanta, Georgia, 2013. Association for Computational Linguistics.
13. *Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky.* Distant supervision for relation extraction without labeled data. pages 1003–1011, 2009.
14. *Valerie Mozharova and Natalia Loukachevitch.* Two-stage approach in russian named entity recognition. In 2016 International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT), pages 1–6. IEEE, 2016.
15. *James Pustejovsky and Amber Stubbs.* Natural language annotation for machine learning. 2013.
16. *Sebastian Riedel, Limin Yao, and Andrew McCallum.* Modeling relations and their mentions without labeled text. In Lecture Notes in Computer Science (including subseries Lecture Notes in Articial Intelligence and Lecture Notes in Bioinformatics), volume 6323 LNAI, pages 148–163, 2010.
17. *A. S. Starostin, V. V. Bocharov, S. V. Alexeeva, A. A. Bodrova, A. S. Chuchunkov, S. S. Dzhumaev, I. V. Emenko, D. V. Granovsky, V. F. Khoroshevsky, I. V. Krylova, M. A. Nikolaeva, I. M. Smurov, and S. Y. Toldova.* FactRuEval 2016: Evaluation of named entity recognition and fact extraction systems for Russian. In Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, pages 702–720, 2016.
18. *Pontus Stenetorp, Sampo Pyysalo, Goran Topic, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii.* brat: a Web-based Tool for NLP-Assisted Text Annotation. In Proceedings of the Demonstrations Session at EACL 2012, Avignon, France, 2012. Association for Computational Linguistics.
19. *Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin.* Attention is all you need. In Advances in Neural Information Processing Systems, volume 2017-Decem, pages 5999–6009, 2017.

20. *Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning.* Position-aware attention and supervised data improve slot lling. In EMNLP 2017 — Conference on Empirical Methods in Natural Language Processing, Proceedings , pages 35–45, 2017.
21. *Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning.* 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, 2017.