004.032.26

Выделение именованных сущностей из текстов распорядительных документов с помощью глубоких нейронных сетей

Березин С.А. (Московский физико-технический институт, Новосибирский государственный университет),

Бондаренко И.Ю. (Новосибирский государственный университет)

Выделение именованных сущностей - это задача извлечения из текстовых данных информации, принадлежащей к заранее определенным категориям, таким как названия организаций, топонимы, имена людей и т.п. В рамках представленной работы был разработан подход, развивающий идеи предшественников по дообучению глубоких нейроных сетей с механизмом внимания архитектуры BERT. Показано, что предварительное обучение языковой модели задачам восстановления маскированного слова и определению семантической связанности двух предложений позволяет заметно улучшить показатели качества решения задачи выделения именованных сущностей. Достигнут один из лучших результатов в задаче выделения именованных сущностей на наборе данных RuREBus, содержащем тексты распорядительных документов министерства экономического развития Российской Федерации. Одной из ключевых особенностей описываемого решения является близость постановки к реальным бизнес-задачам и выделение сущностей не общебытового характера, а специфичных для экономической отрасли.

Ключевые слова: *глубокие нейронные сети, обработка естественного языка, выделение именованных сущностей, BERT*

1. Введение

Выделение именованных сущностей (NER) – это задача выделения в тексте слов или их сочетаний, обозначающих объект или явление определённой категории, например, названия организаций, имена людей и т.д. [1]. Часто выделенные объекты связаны семантическими отношениями (например: «спрос вырос») – обнаружение таких отношения составляет задачу выделения отношений (RE).

Будучи интуитивно понятными для людей, эти задачи долгое время находились за пределами возможностей автоматизированных систем. Многие годы лучшие решения основывались на некоем наборе правил, составленных вручную или автоматически. Значительным прорывом стало использование рекуррентных нейронных сетей [18, 19] и моделей, основанных на векторных представлениях слов, таких как word2vec [2] и GloVe [3]. Тем не менее, качественный скачок в решении этой задачи произошел только с появлением языковых моделей, использующих механизм внимания [4].

Решение поставленной задачи для русского языка существенно усложняется малым числом размеченных наборов данных и, более того, существующие корпуса данных достаточно далеки от типичной бизнес-постановки задачи по следующим причинам:

1) Во-первых, отношения выделены в тексте достаточно плотно (напротив, в бизнес-задачах часто присутствуют лишь 1-2 вхождения отношений на достаточно объемные тексты).

2) Во-вторых, в стандартных корпусах определены отношения бытового и повседневного характера (отношение работы персоны в компании, купли/продажи, владения, родственные отношения, факты рождения и смерти и т. п.), тогда, как в бизнес-задачах обычно требуется выделять отношения, имеющие специфическую природу, связанную с тематикой предметной области.

1.1 Обзор существующих решений

Отправной точкой для исследования задачи выделения именованных сущностей в текстах на русском языке можно считать работу [6], в которой авторы представляют стандартизированный набор данных для обучения алгоритмов выделения именованных сущностей и описывают несколько базовых подходов, послуживших основой для дальнейших работ.

Участники прошедшего в 2016 году соревнования FactRuEval-2016 [14], посвященного задаче выделения именованных сущностей, предложили несколько решений данной проблемы. Так, Сысоев А.А. и Адрианов И.А. предложили подход, основанный на использовании языковой модели word2vec [2] – будучи обученной на задаче предсказания пропущенного слова по окружающим его словам (или предсказания окружающих слов по данному) такая модель способна сформировать семантически значимые векторные представления слов.

В начале 2019 года была опубликована работа [7], в которой авторы описывают подход, основанный на архитектуре CharCNN-BLSTM-CRF, и достигают примечательных результатов.

В том же году, в ходе соревнования BSNLP-2019 [13] выдающиеся результаты были показаны Tsygankova et al., использовавшими подход на основе BiLSTM-CRF [15] с применением эмбеддингов, полученных неизменённой языковой моделью BERT multilingual, и Arkhipov et al., которые имплементировали модифицированную версию архитектуры BERT, путем дообучения мультиязычной архитектуры на текстах на русском языке, скомбинированную с CRF в качестве выходного слоя [16].

2. Описание данных

Для обучения использовался корпус Минэкономразвития, который представляет собой различные отчеты региональных органов о проделанной работе и запланированных мероприятиях, а также прогнозы и планы на будущее. Некоторое подмножество корпуса заранее размечено специальными именованными сущностями (8 классов) и семантическими отношениями на них (11 классов).

По своей постановке задача является задачей классификации слов. В постановке задачи присутствуют (не считая тип 0 – прочие слова) следующие типы выделяемых сущностей:

1) MET (metric) – численный индикатор\показатель, объект, на котором определена операция сравнения (пример: «производительность труда»);
2) ECO (economics) – экономическая сущность (из тех, что не подходят под определение MET) или объект инфраструктуры (пример: «биологических ресурсов»);
3) BIN (binary) – одноразовое действие\бинарная характеристика (есть или нет) (пример: «создание», «строительство»);
4) CMP (compare) – сравнительная характеристика (пример: «выше чем»);
5) QUA (qualitive) – качественная характеристика (пример: «стабильное»);
6) ACT (activity) — принимаемые меры, проводимые мероприятия (пример: «профилактика наркомании»);
7) INST (institutions) — различные учреждения, заведения, структуры и организации (пример: «центр досуга и творчества»);
8) SOC (social) – социальный объект (пример: «кадровая система»)

Разметка корпуса выполнена в формате brat standoff [17] и включает в себя номер объекта или отношения, тип, позицию первого и последнего символов объекта или перечисление аргументов отношения. Пример данных показан на рисунке 1.

```
T127    SOC 6730 6742   правопорядка
T128    BIN 6955 6965   реализации
T129    BIN 7009 7019   достижение
T130    INST 7198 7214  Правительства РК
T131    INST 7387 7403  Правительства РК
R1      GOL Arg1:T3 Arg2:T4
R2      GOL Arg1:T3 Arg2:T5
R3      GOL Arg1:T70 Arg2:T6
R4      GOL Arg1:T70 Arg2:T7
R5      TSK Arg1:T10 Arg2:T11
```

Рисунок 1 – Пример размеченных данных

3. Предобучение языковой модели

Предлагаемые модели для NER основаны на архитектуре BERT [8]. В качестве начальной точки обучения был взят RuBERT [9] — дообученный на текстах русской Википедии и новостных статьях вариант модели BERT. Использование этой, а не мультиязычной модели, как показано далее, даёт заметный прирост в качестве работы. Веса RuBERT использовались в дальнейшем дообучении модели на неразмеченном текстовом корпусе отчетов Минэкономразвития, который описан в предыдущем разделе.

Для обучения модели использовался Google Cloud TPU v2 (tensor processing unit). Размер пакета – 128, скорость обучения – $\mathbf{2 * 10^{-5}}$. Максимальная длина последовательности – 512 токенов, доля маскируемых токенов – 0.15, размер словаря – 119547 токенов. Результаты экспериментов приведены в таблице 1.

Таблица 1 – Результаты экспериментов по обучению BERT

Название модели	Число итераций обучения	Значение loss функции	Результат в задаче NER, F1 macro
BERT Multilingual	0	9.3	0.5458±0.0115
RuBERT	0	9.1	0.555±0.003
BERT Multilingual	210000	2.9	0.430
RuBERT	210000	2.8	0.463

RuBERT	2000	6.2	0.5636
RuBERT	**2500**	5.8	**0.5758**
RuBERT	3000	5.01	0.5732
RuBERT	4000	4.15	0.5332

Как видно в таблице 1, модель довольно быстро переобучается под задачи предсказания маскированного слова и предсказания следующего предложения, что плохо сказывается на её применимости к задаче выделения именованных сущностей.

Лишь спустя несколько экспериментов удалось выявить оптимальное количество итераций обучения. Также, проверялась гипотеза о том, что составление совершенно нового словаря может улучшить показатели, однако, на столь малом объеме текстов модель не смогла хорошо обучиться с нуля, поэтому было принято решение продолжить использование предобученной модели.

Таким образом, была получена модель, адаптированная не только для русского языка в целом, но и конкретно для формального стиля официальных документов.

4. Обучение классификатора

Добавленные к RuBERT слои для решения задачи NER – это классификатор, который принимает векторные представления слов и возвращает категории сущностей, соответствующие этим словам. На вход добавленными слоями принимаются сгенерированные предобученным BERT векторные представления слов, которые затем последовательно принимаются блоком BiLSTM, служащим для выявления зависимостей между токенами предложения. Следующим шагом является полносвязный слой и, наконец, мы используем слой CRF для получения меток классов. Архитектура модели

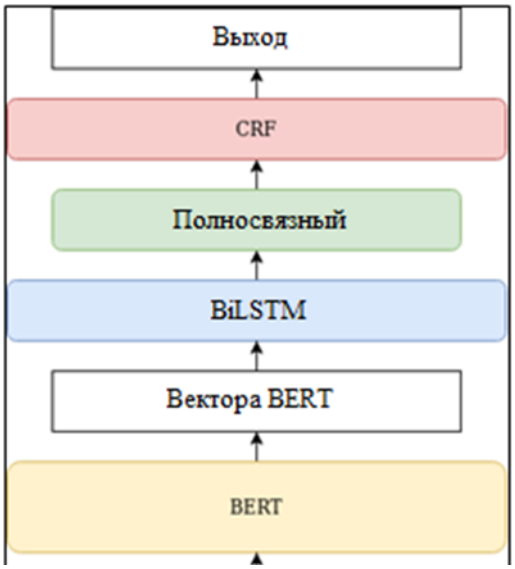
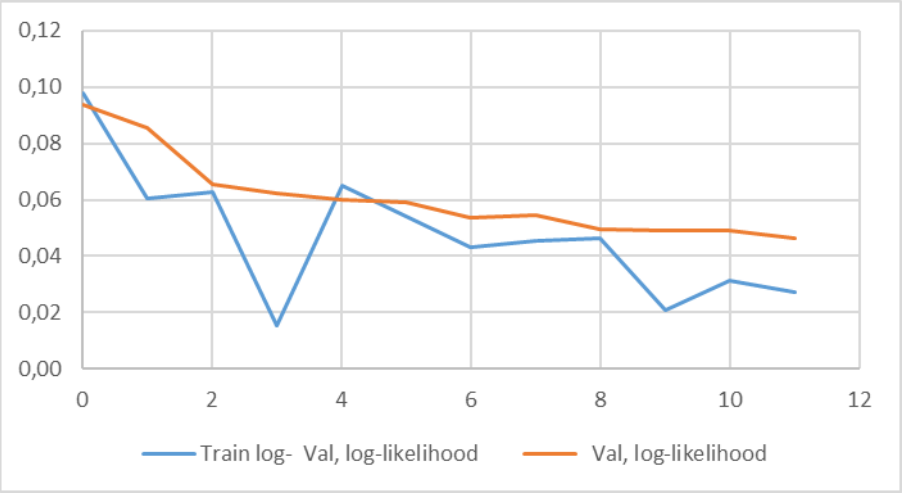


Рисунок 2 – Архитектура нейросети для определения именованных сущностей

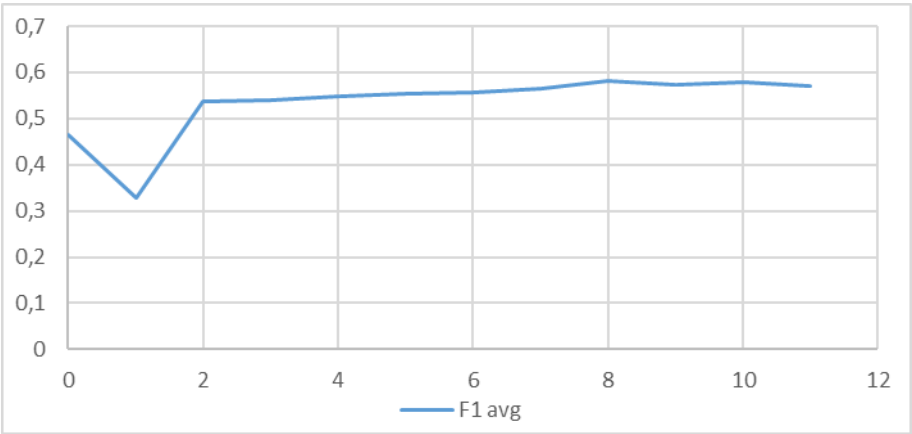


продемонстрирована на рисунке 2.

Рисунок 3 – Значения функции потерь во время обучения

Обучение производилось на GPU Nvidia Tesla P100. Размер мини-выборки составлял 128 экземпляров, скорость обучения равнялась $1 * 10^{-4}$. Для обучения использовалось 32109 примеров, для тестирования 4699 примеров. В качестве меры качества использовалась макро-усреднённая F1-мера. В качестве функции потерь использовалась мультиклассовая кросс-энтропия. Ход эксперимента отображен на рисунках 3-7.

Как видно на рисунке 3, обучение было остановлено раньше, чем в значениях функции потерь на валидационной выборке появились какие-либо признаки переобучения. Причиной этому является неабсолютная корреляция функции потерь с фактической метрикой качества. Если мы обратимся к рисунку 4, то заметим, что F1-мера на валидационной выборке не только вышла на плато, но даже несколько упала, что позволяет говорить о том, что переобучение, де-факто, в некоторой степени уже



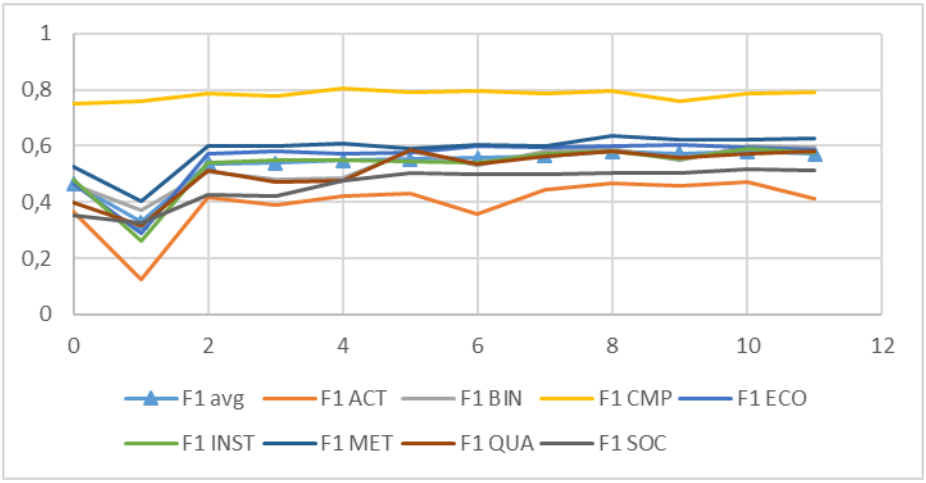


Рисунок 5 – Значения F1-меры во время обучения

произошло.

Рисунок 4 – Значения усреднённой F1-меры во время обучения

Более подробная картина изменения значений F1-меры на валидационных данных продемонстрирована на рисунке 5. Нельзя не обратить внимание на разительно отличающиеся в лучшую сторону показатели качества распознавания сущностей, характеризующих качественные характеристики (QUA). Если ставить задачу выявления только таких сущностей, то F1-мера обученной модели достигает значений вплоть до 0.80.

Также стоит отметить, что, помимо класса QUA, класс социальных объектов (SOC) не оказался подвержен столь значительным флуктуациям, как все остальные классы в начале обучения и стабильно показывал быстрый рост качества вплоть до 5-ой эпохи. Класс принимаемых мероприятий (ACT) стабильно оказывался самым сложным для выделения.

Любопытно также подробнее изучить ход обучения и проанализировать точность и полноту распознавания классов, представленные на рисунках 6 и 7 соответственно.

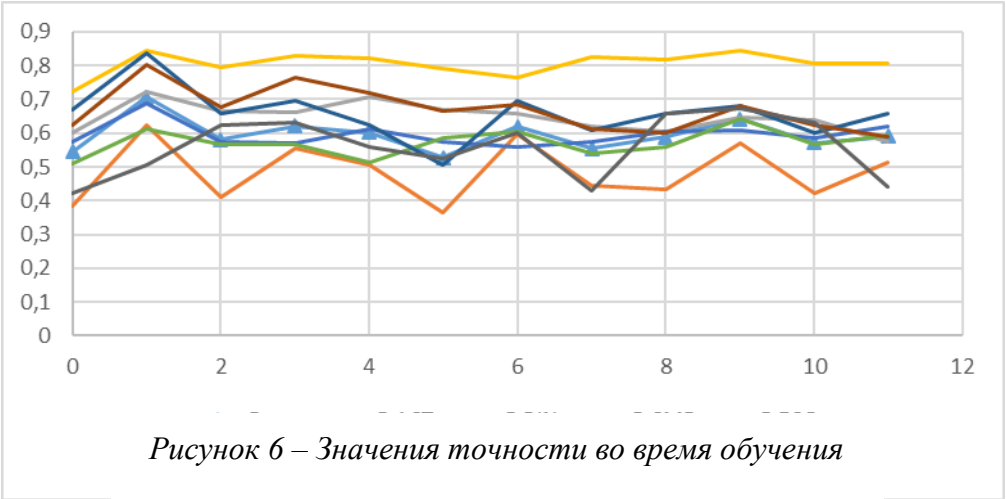


Рисунок 6 – Значения точности во время обучения

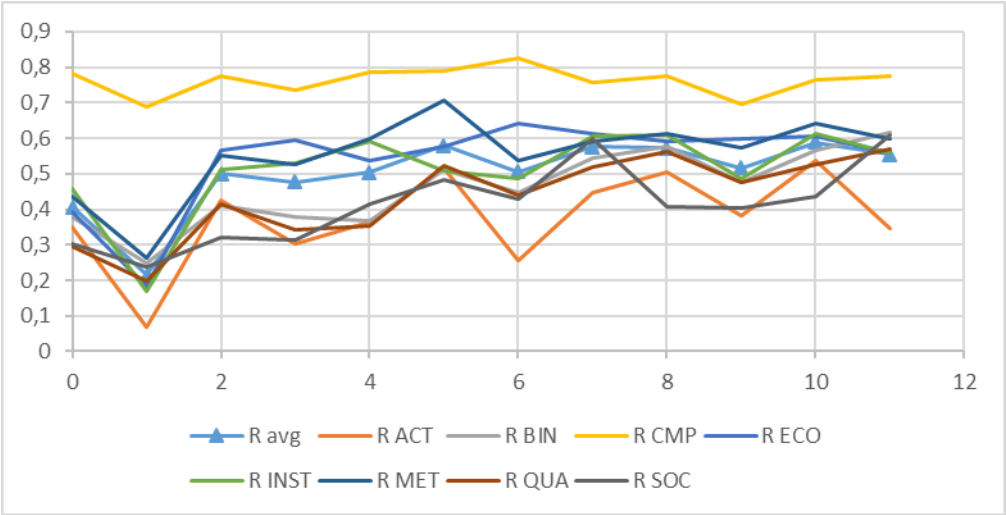


Рисунок 7 – Значения полноты во время обучения

В целом, заметна тенденция модели не отмечать лишних объектов в начале обучения, при этом пропуская много правильных ответов, которая устраняется в дальнейшем.

5. Результаты

В таблице 2 приведено сравнение достигнутых результатов с результатами участников конкурса RuREBus [5], проводимого в рамках конференции Dialog-2020, перед участниками которого ставилась та же задача, что была рассмотрена в ходе данной работы. Важно отметить: несмотря на то, что результат был получен на том же самом тестовом наборе данных, он не был получен во время официальной судейской оценки.

Таблица 2 – Сравнение результатов

Команда	Результат (F1-мера, макро усреднённая)
davletov-aa	0.561
Sdernal	0.464
ksmith	0.463
viby	0.417
dimsolo	0.400
bond005	0.338

Student2020	0.253
Данная работа	**0.576**

6. Заключение

В результате выполненной работы разработано одно из лучших решений задачи выделения именованных сущностей на наборе данных RuREBus.

Продемонстрировано, что предварительное обучение языковой модели задачам восстановления маскированного слова и определению семантической связанности двух предложений позволяет улучшить показатели качества решения задачи выделения именованных сущностей и не требует при этом затрат человеческих усилий на какую-либо разметку и подготовку данных.

В процессе работы над решением были определены следующие пути улучшения:

1) использование более сложных классификаторов – например, байесовской нейронной сети [10];
2) использование более совершенных языковых моделей, таких как ERNIE [11];
3) использование аугментации данных с применением языковой модели для оценки примеров обучения, сгенерированных автоматически из существующих данных [12].

Список литературы

1. Georgios Petasis, Frantz Vichot, Francis Wolinski, Georgios Paliouras Using Machine Learning to Maintain Rule-based Named-Entity Recognition and Classification Systems // ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. 2001. P. 426–433
2. Sysoev A. A., Andrianov I. A. Named Entity Recognition in Russian: the Power of Wiki-Based Approach // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016”: 2016. URL: http://www.dialog-21.ru/media/3433/sysoevaaandrianovia.pdf (дата обращения: 25.09.2020).
3. Pennington Jeffrey, Socher Richard, Manning Christopher Glove: Global Vectors for Word Representation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP): 2014. P. 1532–1543
4. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin Attention is all you need // Proceedings of the 31st International Conference on Neural Information Processing Systems: 2017. P. 6000–6010

5. Ivanin, V., Artemova, E., Batura, T., Ivanov, V., Sarkisyan, V., Tutubalina, E., & Smurov, I. RuREBus-2020 Shared Task: Russian Relation Extraction for Business // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog”. 2020. URL: http://www.dialog-21.ru/media/5098/ivaninvaplusetal-182.pdf (дата обращения: 25.09.2020).
6. Rinat Gareev, Maksim Tkachenko, Valery D Solovyev, Andrey Simanovsky, Vladimir Ivanov Introducing Baselines for Russian Named Entity Recognition // CICLing 2013: Computational Linguistics and Intelligent Text Processing. 2013. P. 329–342
7. Anh Le, Mikhail Burtsev A Deep Neural Network Model for the Task of Named Entity Recognition // International Journal of Machine Learning and Computing vol. 9, no. 1. 2019. URL: http://www.ijmlc.org/vol9/758-ML0025.pdf (дата обращения: 25.09.2020).
8. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // 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). 2019. P. 4171–4186
9. 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. URL: http://www.dialog-21.ru/media/4606/kuratovyplusarkhipovm-025.pdf (дата обращения: 25.09.2020).
10. Vikram Mullachery, Aniruddh Khera, Amir Husain. Bayesian Neural Networks // Digital Culture & Society (DCS): Vol. 4. 2018. URL: https://arxiv.org/abs/1801.07710 (дата обращения: 25.09.2020).
11. Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, Qun Liu ERNIE: Enhanced Language Representation with Informative Entities // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. P. 1441–1451
12. Jonathan Rotsztejn, Nora Hollenstein, Ce Zhang Effectively Combining ETH-DS3Lab at SemEval-2018 Task 7: Recurrent and Convolutional Neural Networks for Relation Classification and Extraction // Proceedings of The 12th International Workshop on Semantic Evaluation. 2018. URL: https://arxiv.org/pdf/1804.02042.pdf (дата обращения: 25.09.2020).
13. Jakub Piskorski et al., The Second Cross-Lingual Challenge on Recognition, Normalization, Classification, and Linking of Named Entities across Slavic Languages // Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing. 2019. P. 63–74
14. 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 // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016”. 2016. P. 688-705

15. Tatiana Tsygankova, Stephen Mayhew, and Dan Roth BSNLP2019 shared task submission: Multisource neural NER transfer // Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing. Association for Computational Linguistics. 2019. P. 75–82
16. Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin Tuning multilingual transformers for language-specific named entity recognition // Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing. 2019. P. 89–93
17. Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou and Jun'ichi Tsujii brat: A Web-based Tool for NLP-Assisted Text Annotation // Proceedings of the Demonstrations Session at EACL 2012. 2012. P. 102–107
18. Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, Chris Dyer Neural Architectures for Named Entity Recognition // Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016. P. 260–270
19. Xuezhe Ma, Eduard Hovy End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016. P. 1064–1074