NEREL: Набор данных на русском языке с вложенными именованными сущностями и отношениями

И. В. Денисов[1] И. С. Рожков[2] Н. В. Лукашевич[3]

NEREL - русский публично доступный набор данных для решения задачи извлечения именованных сущностей и задачи извлечения отношений. Датасет содержит более 56K размеченных сущностей и более 39K отношений. Важным отличием NEREL от предыдущих датасетов является наличие разметки для вложенных именованных сущностей.

Методы извлечения вложенных именованных сущностей отличаются от методов извлечения "плоских" именованных сущностей в первую очередь архитектурой решения. Поскольку NEREL предоставляет аннотации для вложенных сущностей, в работе было проведено сравнение различных подходов к решению этой задачи с переносом на тексты русского языка.

Ключевые слова: извлечение именованных сущностей, извлечение вложенных именованных сущностей, датасет, набор данных.

1. Введение

Большинство наборов данных, размеченных именованными сущностями, содержат упрощенную разметку именованных сущностей, в которой не предполагается, что именованная сущность может быть вложена в другую именованную сущность. Однако такое упрощение приводит к потере информации.

NEREL — новый датасет на русском языке с размеченными именованными сущностями и отношениями между ними (**N**amed **E**ntities and

[1] *Денисов Илья Вячеславович* — аспирант кафедры алгоритмических языков ф-та вычислительной математики и кибернетики МГУ, e-mail: denilv@mail.ru.

Denisov Ilia Viatcheslavovich — graduate student, Lomonosov Moscow State University, Department of Computational Mathematics and Cybernetics, Chair of Algorithmic Languages.

[2] *Рожков Игорь Сергеевич* — студент кафедры алгоритмических языков ф-та вычислительной математики и кибернетики МГУ, e-mail: fulstocky@gmail.com.

Rozhkov Igor Sergeevich — student, Lomonosov Moscow State University, Department of Computational Mathematics and Cybernetics, Chair of Algorithmic Languages.

[3] *Лукашевич Наталья Валентиновна* — ведущий научный сотрудник в Лаборатории анализа информационных ресурсов, Научно-исследовательский вычислительный центр МГУ, e-mail: louk_nat@mail.ru.

Lukashevich Natalia Valentinovna — leading researcher in the Laboratory of Information Resources Analysis, Research Computing Center of Moscow State University.

RELations). Также одна из особенностей NEREL состоит в том, что размечены вложенные именованные сущности и их отношения.

Отношения между сущностями размечаются в рамках связного текста и не ограничиваются уровнем предложения. На Рис. 1 изображен пример вложенных сущностей и отношений между ними, которые связаны с соседними предложениями.

Рис. 1. Предложение имеет вложенные именованные сущности: Мэр Москвы, Москвы, Мэр; Московский драм. театр Ермоловой, Московский, Ермолова.

2. Основные характеристики NEREL

NEREL содержит 29 типов именованных сущностей и 49 типов отношений между сущностями. Размечено более 56K сущностей и 39K отношений в более чем 900 документов Russian Wikinews. Максимальная глубина вложенности сущностей — 6. В Таблице 1 указаны сравнительные характеристики датасетов на русском языке для задачи NER и RE, в некоторых датасетах присутствует информация об отношениях между сущностями.

Датасет	Язык	#NE (Типы)	Макс. глубина	#Rel (Типы)
Gareev [2]	ru	44K (2)	1	–
Collection3 [6]	ru	26.4K(3)	1	–
FactRuEval [10]	ru	12K (3)	2	1K (4)
BSNLP [9]	ru	9K (5)	1	–
RuREBUS [11]	ru	121K (5)	1	14.6K (8)
RURED [3]	ru	22.6K (28)	1	5.3K(34)
NEREL (ours)	ru	56K (29)	6	39K (49)

Таблица 1. Сравнение NEREL и других датасетов на русском языке

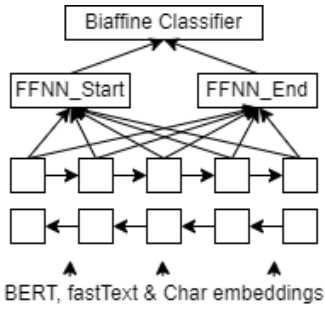

Рис. 2. Архитектура Biaffine NER

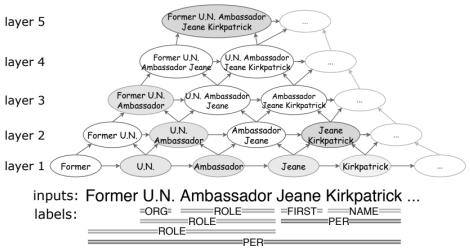

Рис. 3. Архитектура Pyramid NER

3. Методы извлечения вложенных именованных сущностей

Методы извлечения вложенных именованных сущностей архитектурно отличаются от моделей, направленных на извлечение именованных сущностей без вложенности. Для задачи распознавания вложенных именованных сущностей мы рассмотрели несколько моделей глубокого обучения, которые сейчас дают state-of-the-art результаты: Biaffine NER [8], Pyramid NER [7], MRC-NER [4].

3.1. Biaffine NER

В модели Biaffine NER предлагается закодировать входной текст путем объединения векторных представлений полученных из моделей BERT, fastText [5] и символьных представлений. Далее двунаправленный LSTM-слой формирует контекстное представление каждого токена. Эти контекстные представления передаются в FFNN_Start и FFNN_End блоки и формируют разные репрезентации токенов как начала и концов интервалов сущностей (h_s/h_e). На финальном этапе Biaffine Classifier формирует тензор r_m размера $l \times l \times c$ из h_s и h_e. l — число токенов во входном тексте, c — число категорий именованных сущностей + 1 (отсутствие сущности). Тензор r_m содержит в себе оценку вероятности нахождения сущности в каждом возможном интервале с ограничением, что индекс начала сущности i всегда меньше индекса конца сущности ($s_i < e_i$).

3.2. Pyramid NER

В модели Pyramid NER выход нейронной сети при анализе именованных сущностей формируется в виде пирамиды.

Вход модели — это текстовая последовательность, состоящая из T токенов. После кодирования при помощи моделей BERT, fastText и

символьного кодировщика представления рекурсивно подаются в NER-декодирующие слои, генерируя при этом L последовательностей тегов в IOB2-формате (Inside, Outside, Beginning) с длинной $T, T-1, ..., T-L+1$, где L - число декодирующих слоёв. Кодирующие и декодирующие слои связаны между собой, и размер декодирующей последовательности уменьшается путём использования свёрточного слоя с ядром размера 2. Таким образом, на этапе построения модели необходимо чётко понимать максимальную длину сущностей, чтобы поставить соответствующее число декодирующих слоёв.

3.3. MRC-NER

В модели MRC-NER (Machine Reading Comprehension - Машинное Понимание Прочитанного) задача извлечения именованных сущностей ставится как задача ответа на вопросы. Здесь необходимо определить границы ответа **A** на специализированный запрос **Q** в контексте **C**. Набор входных данных формируется из формата NEREL, где каждому слову присвоена метка в BIO-формате (Begin, In, Out). Формат преобразуется в тройки из запроса (Question), ответа (Answer) и контекста (Context). После дальнейшей обработки образуется множество троек $(q_y, x_{start,end}, X)$.

Запросы — заранее определённое множество запросов к каждой категории сущностей. Примеры запросов:

- PERSON: Человек — мужчина, женщина или ребенок.
- ORGANIZATION: Организация — это компания или другая группа людей, которые работают вместе для определенной цели.

Модель в своей основе использует модель BERT, поэтому тройки приводятся к виду: $\{[CLS], q_1, q_2, ..., q_m, [SEP], x_1, x_2, ..., x_n\}$, q_i — токены запроса, x_i — токены контекста. Эта объединённая строка подается на вход BERT-блоку, на выходе получается матрица представления контекст $E \in R^{n \times d}$, где d — размерность последнего слоя BERT, где получено векторное представление запроса.

Для каждого токена вычисляется вероятность того, является ли он началом сущности в запросе. Аналогично проставляется вероятность конца сущности для каждого токена. Возможна ситуация, когда в тексте несколько сущностей, сущности вложены друг в друга или пересекаются, поэтому необходимо определять вероятность того, что конкретная пара start-index и end-index является началом и концом сущности.

4. Эксперименты по извлечению вложенных именованных сущностей

Датасет NEREL был разделен на обучающее, валидационное и тестовое множества — 746/94/93 документов соответственно.

В экспериментах использовалась модель fastText (fT) [5], обученная на текстовых данных Russian Common Crawl[1], и BERT векторные представления текста, полученные на основе модели RuBERT [1].

Метод	P	R	F1
Biaffine, fT	78.84	71.80	75.13
Biaffine, RuBERT	81.92	71.54	76.38
Pyramid, fT	72.70	63.01	67.51
Pyramid, RuBERT	77.73	70.97	74.19
MRC	**85.24**	**84.32**	**84.78**

Таблица 2. Результаты экспериментов по задаче Nested NER

Датасет NEREL позволяет проводить широкий набор экспериментов по задаче Nested NER, т.к. обладает большим количеством типов сущностей и широким набором новостных текстов. На данный момент, наилучшие результаты на тестовой части NEREL показывает модель MRC NER, однако эта модель плохо масштабируется, т.к. к каждому тексту необходимо делать N запросов, где N — число типов сущностей; поэтому второй лучший результат, модель Biaffine NER (RuBert), тоже необходимо рассматривать, если речь идет не только о качестве решения, но и его производительности.

5. Благодарности

Исследование выполнено за счет гранта Российского научного фонда (проект № 20-11-20166).

Список литературы

[1] Yuri Kuratov and Mikhail Arkhipov, “Adaptation of deep bidirectional multilingual transformers for russian language”, *arXiv preprint arXiv:1905.07213*, 2019.

[2] Rinat Gareev et al., “Introducing baselines for russian named entity recognition. In International Conference on Intelligent Text”, *Processing and Computational Linguistics, pages*, 2013, 329–342.

[3] Denis Gordeev et al., “Relation extraction dataset for the russian language.”, *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog” [Komp’iuternaia Lingvistika i Intellektual’nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii “Dialog”]*., 2020.

[1] https://commoncrawl.org/

[4] Xiaoya Li et al., “A unified MRC framework for named entity recognition.”, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.*, 2020, 5849—5859.

[5] Tomas Mikolov et al., “Advances in pre-training distributed word representations.”, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*, 2018.

[6] Valerie Mozharova and Natalia Loukachevitch, “Two-stage approach in russian named entity recognition”, *International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT).*, 2016, 1–6.

[7] Wang Jue et al., “Pyramid: A layered model for nested named entity recognition”, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, 5918—5928.

[8] Juntao Yu et al., “Named entity recognition as dependency parsing”, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, 6470—6476.

[9] Jakub Piskorski et al., “The second crosslingual 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, 64–74.

[10] Anatoly Starostin et al., “Factrueval 2016: Evaluation of named entity recognition and fact extraction systems for russian Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog” [Komp’iuternaia Lingvistika i Intellektual’nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii “Dialog”],”, 2016, 702-720.

[11] Vitaly Ivanin et al., “Rurebus-2020 shared task: Russian relation extraction for business”, *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog” [Komp’iuternaia Lingvistika i Intellektual’nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii “Dialog”]*, 2020.

NEREL: A Russian Dataset with Nested Named Entities and Relations

Denisov I.V., Rozhkov I.S., Lukashevich N.V.

NEREL is a Russian publicly available dataset for solving named entity recognition problem and relation extraction problem. The dataset contains more than 56K tagged entities and more than 39K relationships. An important difference between NEREL and previous datasets is the presence of markup for nested named entities.

The methods of extracting nested named entities differ from the methods of extracting ”flat” named entities primarily by the architecture of the solution. Since NEREL provides annotations for nested entities, the article compared various approaches to solving this problem with the transfer to Russian language domain.

Keywords: NER, nested NER, named entity recognition, nested named entity recognition, dataset.

References

[1] Yuri Kuratov and Mikhail Arkhipov, "Adaptation of deep bidirectional multilingual transformers for russian language", *arXiv preprint arXiv:1905.07213*, 2019.

[2] Rinat Gareev et al., "Introducing baselines for russian named entity recognition. In International Conference on Intelligent Text", *Processing and Computational Linguistics, pages*, 2013, 329-–342.

[3] Denis Gordeev et al., "Relation extraction dataset for the russian language.", *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"]*., 2020.

[4] Xiaoya Li et al., "A unified MRC framework for named entity recognition.", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*., 2020, 5849—5859.

[5] Tomas Mikolov et al., "Advances in pre-training distributed word representations.", *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*., 2018.

[6] Valerie Mozharova and Natalia Loukachevitch, "Two-stage approach in russian named entity recognition", *International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT)*., 2016, 1–6.

[7] Wang Jue et al., "Pyramid: A layered model for nested named entity recognition", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, 5918—5928.

[8] Juntao Yu et al., "Named entity recognition as dependency parsing", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, 6470—6476.

[9] Jakub Piskorski et al., "The second crosslingual 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, 64–74.

[10] Anatoly Starostin et al., "Factrueval 2016: Evaluation of named entity recognition and fact extraction systems for russian Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"],", 2016, 702-720.

[11] Vitaly Ivanin et al., "Rurebus-2020 shared task: Russian relation extraction for business", *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"]*, 2020.