Особенности решения задачи распознавания именованных сущностей на русском датасете

З.Х. Калажоков, Н.А. Андриянов

Кафедра искусственного интеллекта, Финансовый университет при Правительстве Российской Федерации, Москва, Россия

Аннотация: В данной статье рассматриваются особенности реализации моделей для распознавания именованных сущностей. В ходе работы проведен ряд экспериментов как с традиционными моделями, так и с известными архитектурами нейронных сетей, гибридной моделью, рассматриваются особенности результатов, их сравнение и возможные объяснения. В частности, показано, что гибридная модель с добавлением двунаправленной долгой краткосрочной памяти может давать более качественные результаты, чем базовая двунаправленная модель представлений на основе трансформеров. Также показано, что улучшенная путем добавления прореживающего слоя для регуляризации, взвешенной функции потерь и линейного классификатора поверх выходов, двунаправленная модель представлений на основе трансформеров может давать высокие значения метрик. Для наглядности в работе приведены графики обучения моделей и таблицы с метриками для сравнения. В процессе работы сформированы выводы и рекомендации.

Ключевые слова: анализ текста, искусственный интеллект, распознавание именованных сущностей, нейронные сети, глубокое обучение, машинное обучение.

Введение

Задача распознавания именованных сущностей [1] является одной из основных задач в области обработки естественного языка. Иногда она рассматривается как составная часть других задач обработки естественного языка, но часто выступает и как самостоятельная задача. Задача заключается в распознавании и классификации определенных частей неструктурированного текста, известных как "именованные сущности", которые обычно представляют собой объекты реального мира: людей, названия организаций, географические названия, даты, счета, денежные суммы, реквизиты, обозначения терминов разных предметных областей и многое другое. Задачу распознавания именованных сущностей применяют не всегда, так как не каждая задача требует этого, но в подходящих сценариях -

это мощный инструмент. Активно используется в системах распознавания накладных, счетов, квитанций в сочетании с другими инструментами.

В поисках лучшего подхода исследователи экспериментируют с разными технологиями и исходными данными, вследствие чего существует ряд методов решения данной задачи, которые характеризуются различным качеством результатов и возникает необходимость выбора наиболее подходящего подхода для конкретной задачи. В [1,2] авторы приводят анализ существующих методов с точки зрения отдельных категорий задач, выделяют несколько категорий задач и для каждой обсуждаются качество решения, особенности и проблемы методов.

В этой статье приводятся результаты экспериментов обучения разных моделей распознавания именованных сущностей на датасете для русского языка, описание решения возникающих проблем и сравнение результатов для нескольких выбранных методов. Эксперименты проводились с целью изучения и подбора наилучших моделей для дальнейшего решения задачи распознавания именованных сущностей на данных, которые будут извлечены из фото бланков и накладных, в рамках решения задачи распознавания документов.

Основная часть

1. Традиционный подход.

Для начала рассматривалась модель условных случайных полей (Conditional Random Fields - CRF) [3], которая представляет собой вероятностную модель, в которой многие принципы унаследованы от скрытой марковской модели. CRF отличается от скрытой марковской модели главным образом тем, что наблюдаемые значения зависят не только от

текущего скрытого состояния, но от всех скрытых состояний сразу, что дает возможность захвата контекста. Это нейронная сеть с двумя полносвязными слоями, которые используются независимо друг от друга, один слой характеризует матрицу переходов, другой - матрицу наблюдений. CRF является относительно простой и легковесной моделью, может уступать в точности современным моделям, но может быть достаточным для большого круга задач.

Была обучена модель на датасете для русского языка factRuEval-2016. Результаты приведены в Таблице 1.

Таблица № 1

Результаты классификации для модели CRF

	Точность	Полнота	f1-мера
Доля верных ответов			0.98
Макросреднее	0.93	0.90	0.91
Взвешенное среднее	0.98	0.98	0.98

Полученные основные метрики, на который стоит обратить внимание, это F1-Мера, представляющее собой гармоническое среднее точности и полноты, а также точность и полнота. Модель дает лучший результат для класса "не сущности", высокие результаты для имен людей, локаций, испытывает проблемы с многословными названиями организаций, чуть лучше определяет длинные названия локаций и людей. CRF-модель показывает довольно высокие результаты, доля верных ответов = 0.98, но это из-за доминирования класса "не сущности", поэтому лучше обратить внимание на F1-меру, среднее по классам без учета весов (макросреднее) = 0.91, показывает хороший баланс между точностью и полнотой. Можно сделать вывод, что такая модель хорошо работает для простых сущностей (имена и локации), но с вложенными токенами испытывает трудности.

Улучшить данную модель можно рядом способов: добавить морфологические признаки в CRF [4], что помогает различать омонимы; для финансовых документов может оказаться полезным добавление специальных специфических признаков как ИНН, суммы, даты, которые встречаются в подобного рода документах; также полезно применить постобработку правилами, чтобы исключить случаи ошибок CRF нахождения многословных названий организаций; оптимизация гиперпараметров влияет на точность и переобучение, их регулирование поможет в тонкой настройке модели.

В качестве первого улучшения сети был рассмотрен вариант расширения признаков. Результаты приведены в Таблице 2.

Таблица № 2

Результаты классификации для модели CRF с расширенными признаками

	Точность	Полнота	f1-мера
Доля верных ответов			0.98
Макросреднее	0.93	0.90	0.92
Взвешенное среднее	0.98	0.98	0.98

В эксперименте наблюдается незначительный рост F1-меры для имен людей, некоторое ухудшение для локаций и организаций, возможно, из-за шума в новых признаках или переобучения. В итоге имеем минимальный прирост качества для указанных сущностей.

В качестве еще одного эксперимента по улучшению была рассмотрена модель CRF в комбинации с правилами. Под правилами имеются в виду некоторые заданные установки, например, если токен состоит из более трех заглавных букв, то он классифицируется как организация и др. Результаты этого эксперимента представлены в Таблице 3:

Таблица № 3

Результаты классификации для модели CRF в комбинации с правилами

	Точность	Полнота	f1-мера
Доля верных ответов			0.86
Макросреднее	0.79	0.52	0.50
Взвешенное среднее	0.94	0.86	0.87

Эти результаты показывают сильный дисбаланс между точностью и полнотой: Для начального токена в многословном имени: полнота=1.00 (находит все имена), но точность=0.25 (75% предсказаний — ложные срабатывания). Для внутреннего токена: полнота=0.01 (почти не находит), но точность=0.93 (когда находит - почти всегда верно). В результате наблюдается падение качества модели - низкая полнота для начального токена в локации (0.14) и внутреннего токена имени (0.01), что возможно объясняется слишком строгими правилами; при этом для начального токена имени наблюдается полнота = 1.00, но точность = 0.25, что указывает на то, что правила добавляют ложные срабатывания. В качестве рекомендации можно указать то, что возможным решением будет смягчение правил - это может улучшить качество модели.

2. Нейросетевой подход

Были обучены нейросетевые модели. Конечно такие модели требуют для обучения графический процессор, в отличие от ранее рассмотренных моделей. Результаты для модели двунаправленной долгой краткосрочной памяти (Bidirectional Long Short-Term Memory - BiLSTM)+CRF [5], базового БЕРТ (Bidirectional Encoder Representations from Transformers - BERT) [6], гибридной модели BERT+BiLSTM представлены в таблицах. В [7] авторы приводят результат гибридной модели BiLSTM+CRF, который показывает улучшение метрик по сравнению с традиционными CRF и методом опорных векторов.

Таблица № 4

Результаты классификации для модели CRF в комбинации с правилами

	Точность	Полнота	f1-мера
Макросреднее	0.94	0.95	0.95
Взвешенное среднее	0.94	0.95	0.95

Результаты эксперимента выглядят хорошо, F1-мера = 0.95, что выше, чем в базовом CRF, имена и локации распознаются почти идеально. Организации - сложный класс, но улучшения также очевидны. BiLSTM учитывает контекст, это особенно полезно для русского текста.

Далее были рассмотрены базовая модель BERT и гибридная модели с гипотезой, что связка BERT+BiLSTM даст улучшение базовой модели. Результаты для 10-й эпохи приведены в Таблице 5.

Таблица № 5

Результаты классификации для базовой модели BERT и гибридной с добавлением BiLSTM

Модель	f1-мера	Точность	Полнота
BERT	0.614938	0.539066	0.715665
BERT+BiLSTM	0.742307	0.725863	0.759513

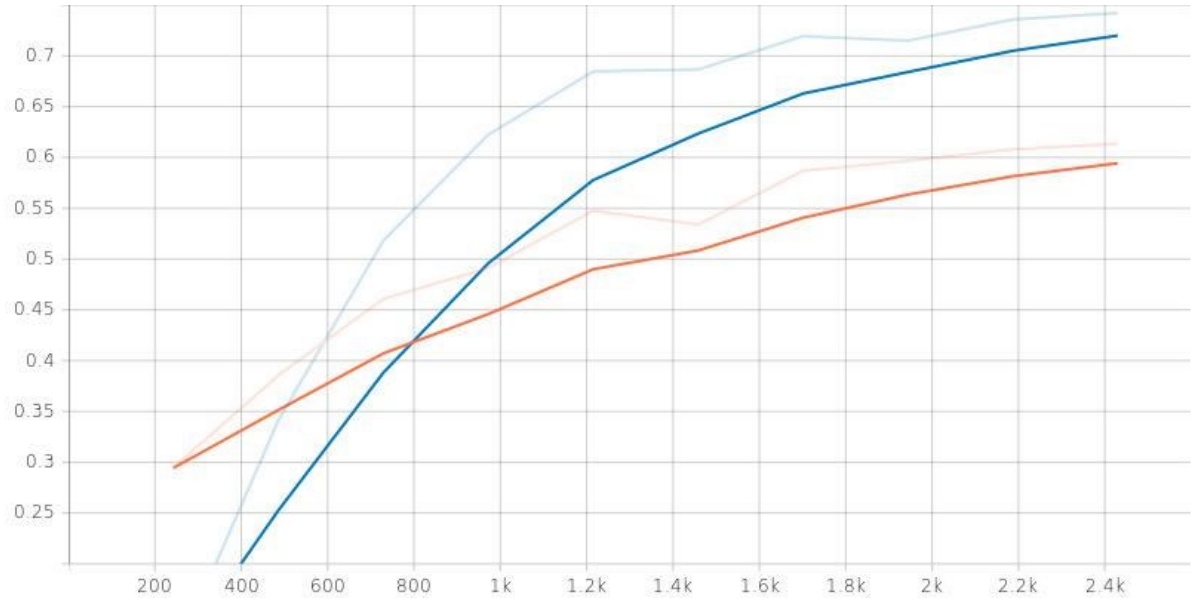

Рис. 1. – Метрика F1-мера для базовой модели BERT и гибридной с добавлением BiLSTM

Действительно, гибридная модель показала улучшение базовой на 12.7% по F1-score, что подтверждает гипотезу о пользе BiLSTM для задач NER. Обучение базовой модели стабильное, но метрики растут медленнее после 7-й эпохи. Гибридная модель показала быстрый рост качества после 3-й эпохи - резкий скачок F1, стабильное улучшение до конца обучения. Интерпретировать улучшение модели можно предположением, что слой BiLSTM позволяет модели лучше учитывать последовательные зависимости между токенами, что для задачи распознавания именованных сущностей критично. Использование взвешенных потерь смещают фокус модели на сущности, а не преобладающий класс "не сущность", что объясняет высокие значения полноты.

Рассматривалась модель BERT[6] с улучшениями. Базовый моделью является RuBERT-base-cased от DeepPavlov. Архитектура дополнена слоем прореживания для регуляризации, а также поверх выходов BERT добавлен линейный классификатор. Также используется взвешенная функция потерь

(метка "не сущность" имеет вес 1, остальные метки имеют вес 3, что призвано решать вопрос дисбаланса классов, т.к. класса "не сущность" больше остальных именованных сущностей). Модель дала на 5-й эпохе следующие результаты:

Таблица № 6

Результаты классификации для модели BERT с улучшениями

f1-мера	Точность	Полнота
0.981667	0.980721	0.982615

Выводы

В качестве выводов можно выделить несколько особенностей, выявленных в рассмотренных моделях для задачи NER.

1. CRF [3] - это хороший базовый метод, но на сложных классах как внутренний токен организации достигает потолка.
2. Правила могут как улучшить модель так и навредить ей, нужно адаптировать их к данным, настроить. Хороший вариант - CRF с фичами [4] вкупе с удачно подобранными правилами. Для задач анализа документов работа с фичами представляется полезной.
3. BiLSTM + CRF [5] - хороший вариант, сочетающий контекстную обработку (BiLSTM) и структурные ограничения (CRF).
4. В сравнении с базовой BERT хорошие результаты показала гибридная модель BERT+BiLSTM, что согласуется с выводами других исследований [8, 9].
5. Модель BERT с улучшениями, как описано в эксперименте, дала самый высокий результат среди рассмотренных.
6. Рассмотренные приемы можно применять на пользовательском наборе данных, возможно комбинируя между собой, как это показано в гибридной модели. Имеет смысл пробовать построить

гибридную модель BERT+BiLSTM, но не с базовой BERT, а с улучшениями. Также при дисбалансе классов можно пробовать применить фокальные потери, вместо взвешенных весов, но за счет интерпретируемости.

7. В ходе работы реализован ряд моделей и рассмотрены результаты на русском датасете, что представляется полезным в свете множества работ для англоязычных датасетов и относительного меньшинства для русскоязычных датасетов.

Статья подготовлена по результатам исследований, выполненных за счет бюджетных средств по государственному заданию Финуниверситета.

Литература

1. Lagutina N.S., Vasilyev A.M., Zafievsky D.D. Name Entity Recognition Tasks: Technologies and Tools. // Modeling and analysis of information systems. 2023. Vol. 30, no. 1. Pp. 64-85.
2. Anh L.T., Arkhipov M.Y., Burtsev M.S. Application of a Hybrid Bi-LSTM-CRF model to the task of Russian Named Entity Recognition. // AIST. 2020. Pp. 91–103.
3. Sutton C., McCallum A. An Introduction to Conditional Random Fields. // Foundations and Trends in Machine Learning. 2012. DOI: 10.1561/2200000013.
4. Tkachenko M., Simanovsky A. Named Entity Recognition: Exploring Features. // KONVENS. 2012. Pp. 118-127.
5. Huang Z., Xu W., Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging // arXiv.org. – 2015. – URL: arxiv.org/abs/1508.01991 (дата обращения: 21.07.2025).

6. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. // NAACL-HLT. 2019. Pp. 4171-4186.
7. Lample G., Ballesteros M., Subramanian S., Kawakami K., Dyer C. Neural Architectures for Named Entity Recognition. // NAACL-HLT. 2016. Pp. 260-270.
8. Peters M., et al. Deep Contextualized Word Representations. // NAACL-HLT. 2018. Pp. 2227–2237.
9. Kuratov Y., Arkhipov M. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language // arXiv. – 2019. – URL: arxiv.org/abs/1905.07213 (дата обращения: 21.07.2025).
10. Kondratyuk D., Straka M. 75 Languages, 1 Model: Parsing Universal Dependencies Universally. // EMNLP. 2019.

References

1. Lagutina N.S., Vasilyev A.M., Zafievsky D.D. Modeling and analysis of information systems. 2023. Vol. 30, no. 1. Pp. 64-85.
2. Anh L.T., Arkhipov M.Y., Burtsev M.S. AIST. 2020. Pp. 91–103.
3. Sutton C., McCallum A. Foundations and Trends in Machine Learning. 2012. DOI: 10.1561/2200000013.
4. Tkachenko M., Simanovsky A. KONVENS. 2012. Pp. 118-127.
5. Huang Z., Xu W., Yu K. arXiv: 1508.01991. 2015. URL: org/abs/1508.01991 (date assessed: 21.07.2025).
6. Devlin J., Chang M.-W., Lee K., Toutanova K. NAACL-HLT. 2019. Pp. 4171-4186.
7. Lample G., Ballesteros M., Subramanian S., Kawakami K., Dyer C. NAACL-HLT. 2016. Pp. 260-270.
8. Peters M., et al. NAACL-HLT. 2018. Pp. 2227–2237.

9. Kuratov Y., Arkhipov M. arXiv: 1905.07213. 2019. URL: arxiv.org/abs/1905.07213 (date assessed: 21.07.2025).
10. Kondratyuk D., Straka M. 75 Languages, 1 Model: Parsing Universal Dependencies Universally. EMNLP. 2019.

Дата поступления: 13.07.2025

Дата публикации: 25.08.2025