ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

2022 Управление, вычислительная техника и информатика № 58

Tomsk State University Journal of Control and Computer Science

Научная статья
УДК 519.2
doi: 10.17223/19988605/58/7

Подход к распознаванию именованных сущностей на примере технологических терминов в условиях ограниченной обучающей выборки

Алексей Дмитриевич Кульневич[1], Александр Алексеевич Кошечкин[2], Святослав Васильевич Карев[3], Александр Владимирович Замятин[4]

[1, 2, 3, 4] *Томский государственный университет, Томск, Россия*
[1] *kulnevich94@mail.ru*
[2] *kaa1994g@mail.ru*
[3] *svyatoslav.karev@live.ru*
[4] *avzamyatin@inbox.ru*

Аннотация. Предлагается подход к распознаванию именованных сущностей применительно к области технологических терминов в условиях ограниченной выборки с применением предварительно обученных нейронных сетей в комбинации со статистической моделью. Исследуется применимость каждой модели в отдельности и их комбинация. Обосновывается применимость предложенного подхода для ограниченной обучающей выборки.

Ключевые слова: распознавание технологических терминов; распознавание именованных сущностей; комбинирование моделей; Bi-LSTM (bidirectional long short-term memory); CRF (conditional random field)

Для цитирования: Кульневич А.Д., Кошечкин А.А., Карев С. В., Замятин А.В. Подход к распознаванию именованных сущностей на примере технологических терминов в условиях ограниченной обучающей выборки // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2022. № 58. С. 71–81. doi: 10.17223/19988605/58/7

Original article
doi: 10.17223/19988605/58/7

An approach to recognizing named entities using the example of technological terms in a limited training sample

Alexey D. Kulnevich[1], Alexander A. Koshechkin[2], Svyatoslav V. Karev[3], Alexander V. Zamyatin[4]

[1, 2, 3, 4] *Tomsk State University, Tomsk, Russian Federation*
[1] *kulnevich94@mail.ru*
[2] *kaa1994g@mail.ru*
[3] *svyatoslav.karev@live.ru*
[4] *avzamyatin@inbox.ru*

Abstract. The paper considers the problem of recognizing named entities by the example of technological terms, a named entity is a word or phrase denoting an object or phenomena of a certain category. Automatic recognition of technological terms allows companies to optimize business processes. Recognizing named entities for a limited training sample is a non-trivial task.

Currently, the standard for recognizing named entities are conditional random field methods (conditional random field, CRF) and bidirectional long-term short-term memory network (bidirectional long-term short-term memory, Bi-LSTM).

The paper proposes an approach that is a combination of a statistical (CRF) and a neural network (Bi-SM-CRF) model. The main advantage of using the CRF model is a slight increase in training time against the background of providing additional information for the subsequent Bi-LSTM-CRF model, which will allow you to learn more effectively in a limited sample. Two approaches are used to convert text to feature space: extracting the syntactic properties of words for a statistical model and converting text to a vector using the Sci-Bert language model.

Within the framework of the work, a significant improvement in the quality of recognition of technological terms was demonstrated due to the combination of statistical and neural network models of machine learning and the use of a domain-oriented language model for vector representation of scientific texts. This made it possible to improve the quality of recognition of technological terms using the f1-score metric by 12% when training on 800 texts compared to the traditional approach.

Keywords: technology term recognition; named entity recognition; model combination; Bi-LSTM (bidirectional long short-term memory); CRF (conditional random field)

***For citation*:** Kulnevich, A.D., Koshechkin, A.A., Karev, S.V., Zamyatin, A.V. An approach to recognizing named entities using the example of technological terms in a limited training sample. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitelnaja tehnika i informatika – Tomsk State University Journal of Control and Computer Science*. 58. pp. 71–81. doi: 10.17223/19988605/58/7

Именованная сущность – это термин, обозначающий предмет или явление определенной категории. Распознавание именованных сущностей – задача поиска в текстах именованных сущностей определенной категории для последующего анализа [1]. Одним из примеров такой категории являются технологические термины.

На сегодняшний день в научных журналах публикуется значительное число статей, описывающих новые технологические тренды, позволяющие решать передовые задачи в различных отраслях промышленности. Количество материала, требующего изучения, уже давно превосходит возможности человека. В связи с этим автоматическое распознавание технологических терминов в таких статьях представляет интерес для индустрии, поскольку их своевременное выявление позволит компаниям быстро принимать решения по оптимизации внутренних бизнес-процессов и корректировке приоритетов. При этом на данный момент существуют многочисленные сложности по распознаванию технологических терминов. Например, такой термин может представлять собой аббревиатуру, несколько слов, терминов-словосочетаний, при этом ситуация осложняется и тем, что слова могут размещаться в различных частях одного предложения.

На данный момент задача распознавания именованных сущностей успешно решается для таких категорий, как персоны, локации и организации, поскольку их контекст достаточно однообразен [2]. Задачи распознавания именованных сущностей, относящихся к специализированным категориям, решаются при аналогичном объеме обучающей выборки с более низким качеством. Однако увеличение размеров обучающей выборки с целью увеличения качества распознавания не всегда возможно в силу необходимости предметной экспертизы для разметки данных [3].

Сегодня наиболее популярными и эффективными методами для решения задачи распознавания именованных сущностей являются метод условных случайных полей CRF [4] и двунаправленная сеть долгосрочной краткосрочной памяти Bi-LSTM [5, 6].

В работе [7] представлено успешное применение метода CRF для распознавания именованных сущностей на примере технологических терминов. Авторами предложен следующий каскадный подход: проводится классификация предложений в анализируемых текстах на предмет наличия в них технологических терминов с последующим распознаванием технологических терминов внутри классифицированных предложений. Используемый набор данных содержал 240 тыс. предложений, из которых 10 тыс. предложений содержали различные технологические термины. Оценка качества классификации предложений на предмет присутствия технологических терминов и их распознавания внутри предложений по метрике f1-score достигла 93 и 96%, соответственно.

В работе [8] представлены результаты применения Bi-LSTM в задачах распознавания именованных сущностей. Для обучения модели использовались признаки GloVe [9] для слов, признаки символов, извлекаемые с помощью CNN, а также наличие заглавной буквы в начале слова и то, является ли слово лексиконом. Использовалось два набора данных: CoNLL-2003 и OntoNotes 5.0/CoNLL-2012,

содержащих 23 499 и 81 828 именованных сущностей соответственно. Оценка качества распознавания именованных сущностей по метрике f1-score составила 94,03 (± 0,23) и 84,57% (± 0,27) для каждого набора данных соответственно.

В свою очередь, перед применением Bi-LSTM, как и любой другой нейронной сети, к текстам необходимо провести операцию векторизации [10, 11]. На данный момент использование языковых моделей, таких как Bert (Bidirectional encoder representations from transformers), является базовым подходом в данной области в силу особенности преобразования входных данных: каждое слово имеет разное векторное представление в зависимости от контекста предложения, в то время как при использовании word2vec векторное представление каждого слова является фиксированным [12]. Однако для векторизации научных публикаций предпочтительнее использовать модель Sci-Bert, являющуюся языковой моделью Bert, дообученной на большом массиве научных текстов [13].

Стандартные языковые модели предполагают использование для векторизации текста весов только последнего слоя сети. Однако в работе [14] показано, что использование весов нескольких последних слоев сети благоприятно сказывается на конечном результате в задаче распознавания именованных сущностей. В работе [15] для повышения качества извлечения именованных сущностей предлагают использовать технику Scalar mix (Scalar Mixing Weights). Ее смысл заключается в вычислении среднего арифметического по нескольким скрытым слоям и использовании его при вычислении результатов работы языковой модели.

Помимо правильного преобразования слов в вектор, для задачи распознавания именованных сущностей важно правильно подобрать модель. Одним из популярных и наиболее качественных подходов для задачи распознавания именованных сущностей является объединение моделей CRF и Bi-LSTM (Bi-LSTM-CRF) [16]. Предложенный авторами гибридный подход заключается в использовании метода CRF вместо стандартной функции активации на последнем слое Bi-LSTM. Это привело к увеличению качества распознавания именованных сущностей на 3% по метрике f1-score по сравнению с Bi-LSTM, показавшей результаты на уровне 81% для набора данных CoNLL-2003.

Основной проблемой описанных выше работ является необходимость использования больших объемов данных для обучения применяемых моделей, что, в свою очередь, требует как временных, так и высококвалифицированных экспертных ресурсов для разметки данных. Особенно остро данная проблема проявляется в случае поиска технологических терминов, так как для точной разметки выборки необходимы экспертные знания в различных предметных областях. В связи с этим в данной работе предлагается подход к распознаванию именованных сущностей в условиях обучающей выборки ограниченного размера, который позволит быстрее сходиться к приемлемому качеству, не требуя огромных размеченных выборок.

1. Распознавание на основе комбинирования статистической (CRF) и нейросетевой (Bi-LSTM-CRF) моделей

Ситуация в рамках задачи распознавания технологических терминов является стандартной для моделей глубокого обучения – недостаток размеченных данных не позволяет решить ее существующими методами, так как они требуют больших объемов данных. В связи с этим предлагается подход, представляющий собой комбинирование статистической (CRF) и нейросетевой (Bi-LSTM-CRF) моделей. Основным преимуществом привлечения CRF-модели является незначительное увеличение времени обучения на фоне предоставления для последующей Bi-LSTM-CRF-модели дополнительной информации, что позволит эффективнее обучаться в условиях ограниченной выборки.

На рис. 1 представлен подход, являющийся базовым на практике (*a*), и подход, предложенный в данной работе (*b*).

Входными данными для предложенного подхода являются аннотации научных статей. Для преобразования текста в признаковое пространство на следующем шаге используется два подхода – извлечение синтаксических свойств слов для статистической модели и преобразование текста в вектор с помощью языковой модели Sci-Bert.

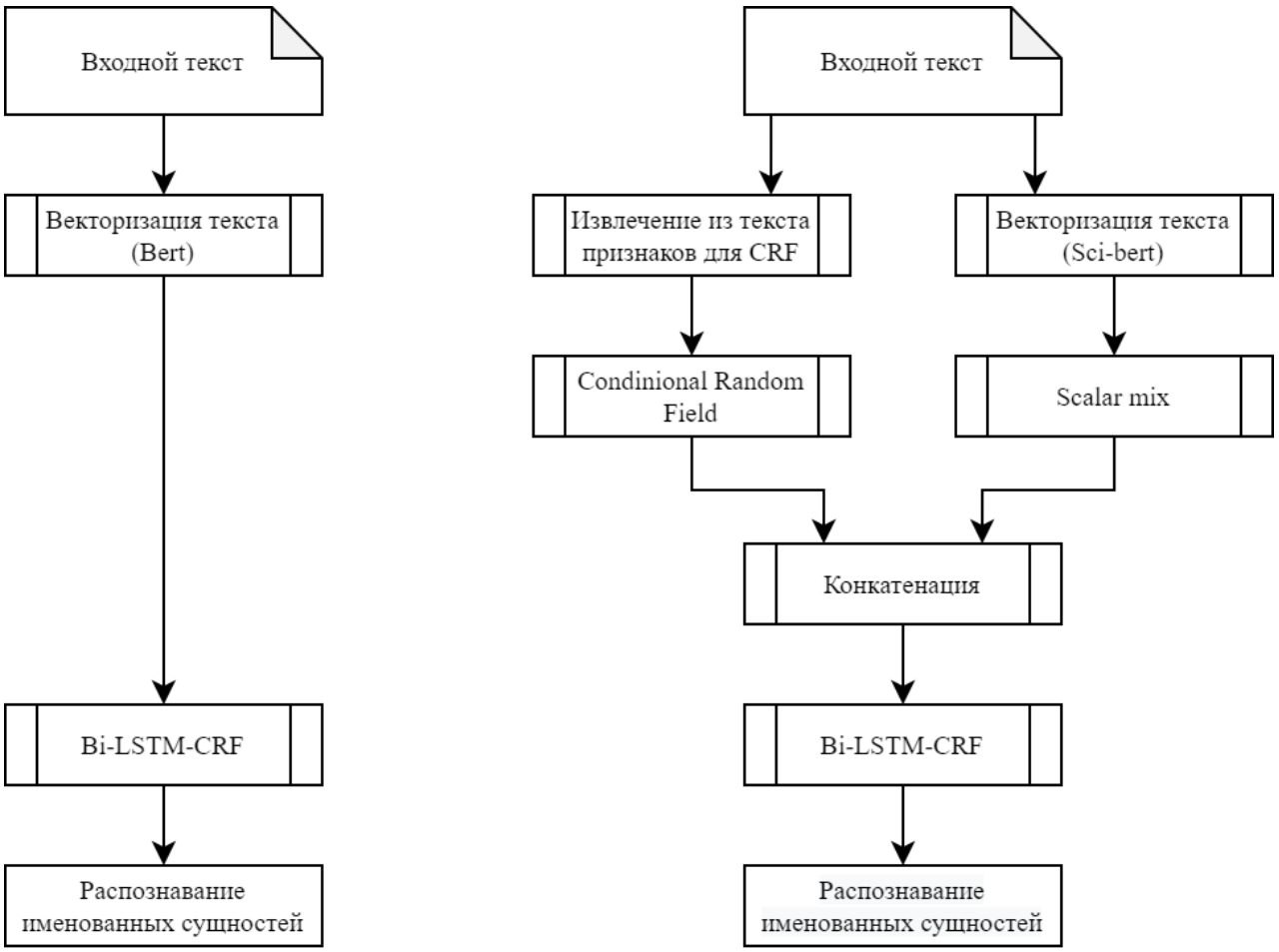

Рис. 1. Подходы к распознаванию именованных сущностей
Fig. 1. Approaches to recognizing named entities

В основе Sci-Bert лежит языковая модель BERT, основанная на трансформере – архитектуре глубоких нейронных сетей с механизмом внимания, изучающим взаимоотношения между словами в тексте [17]. Как правило, трансформер включает две части – энкодер и декодер, но, поскольку исходная задача BERT – восстанавливать пропущенные слова, для задачи векторного представления слов требуется только энкодер.

В противоположность направленным архитектурам, которые считывают текст последовательно слева направо или наоборот, трансформер считывает последовательность слов за один раз. Таким образом, можно сказать, что это ненаправленная модель, позволяющая изучить контекст сразу как слева, так и справа.

При обучении BERT решает две задачи:

1. Создание маскированной языковой модели.

2. Предсказание следующего предложения.

Создание маскированной языковой модели. Перед подачей последовательности слов в BERT 15% слов заменяется маскирующим токеном, и модель пытается их предсказать (рис. 2). Технически это происходит следующим образом:

1. Добавление слоя классификации после энкодера.

2. Маскированные слова помечаются специальным тэгом, который имеет специальное обозначения для токенизатора.

3. Добавление слоя для классификации (состоит, как правило, из одного или нескольких скрытых слоев и softmax-функции активации).

Рис. 2. Маскированная языковая модель
Fig. 2. Masked language model

Предсказание следующего предложения. Для разметки моделью, где начинается следующее предложение, текстовый вход обрабатывается следующим образом (рис. 3):

1. CLS-токен вставляется в начало первого предложения и SEP-токен вставляется в конец каждого.

2. Вектор предложения, обозначающий предложение *A* или предложение *B*, добавляется к каждому токену.

3. Вектор позиции слова складывается с каждым токеном, показывая, где его место в предложении относительно других.

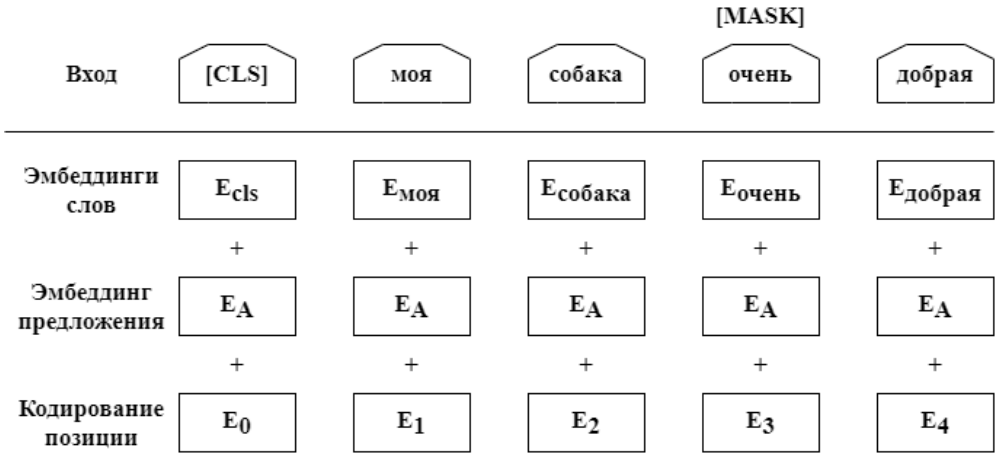

Рис. 3. Предсказание недостающих слов с помощью BERT
Fig. 3. Predicting missing words using BERT

Сам процесс предсказания выглядит следующим образом:

1. Предложение проходит через трансформер-модель.

2. Выход CLS-токена трансформируется в вектор размерности 2 × 1, используя скрытый слой.

3. Результирующий вектор передается в softmax-функцию активации.

Для увеличения качества распознавания из модели Sci-Bert извлекается не только вектор последнего слоя, но и несколько последних скрытых слоев. После извлечения их значения усредняются (Scalar

Mix), чтобы информация сохранилась в векторе и при этом размерность не изменилась по сравнению со стандартным извлечением последнего слоя. В сумме это дает следующее преимущество: нейросетевая модель получает на вход оптимизированное признаковое пространство по сравнению с традиционным подходом, что благоприятно влияет на качество распознавания именованных сущностей.

После извлечения признаков из текста синтаксические свойства слов подаются в модель CRF, а ее предсказания используются как признак для конечной модели. Предсказания CRF-модели и векторное представление текста, сформированное с помощью Sci-Bert с применением Scalar Mix, объединяются с помощью операции конкатенации, и ее результат поступает на вход модели Bi-LSTM-CRF. Выходом модели Bi-LSTM-CRF является размеченная последовательность с информации о принадлежности каждого слова к классу технологических терминов.

2. Вычислительный эксперимент

Для проверки работоспособности и оценки эффективности предложенного метода для задачи распознавания именованных сущностей необходимо сравнить предложенный метод с аналогами (CRF, Bi-LSTM-CRF), а также продемонстрировать оправданность отдельной модификации исходной модели (Bi-LSTM-CRF с применением Sci-Bert-Scalar mix).

В экспериментах с Bi-LSTM использовались подобранные оптимальные гиперпараметры:

– размерность Sci-Bert embeddings: 768;

– количество слоев: 2;

– размерность LSTM-ячеек слоев: 64;

– dropout: 0.2;

– оптимизатор: Adam.

В экспериментах CRF использовались следующие признаки, стандартные для данной модели:

– регистр (нижний, верхний);

– начало с заглавной буквы;

– заголовок;

– число;

– форма слова;

– является ли токен стоп-словом;

– частеречная разметка.

При проведении эксперимента необходимо соблюсти баланс между временем вычислений и объективностью оценки. Стандартным методом формирования обучающей и тестовой выборок является случайное разбиение в заданной пропорции, которое проводится несколько раз с последующим усреднением результатов. Данный метод идеально подходит для ситуаций, когда нужно оценить эффективность различных моделей при изменяющемся объеме обучающей выборки. Для рассмотрения работы моделей в условиях обучающей выборки различного размера обучение проводилось на 1/5, 2/5, 3/5 и 4/5 от всех текстов, а тестирование – на оставшихся. На каждой стадии обучающая и тестовая выборки формировались 10 раз случайным образом с соблюдением соответствующих пропорций с последующим усреднением результатов. В качестве метрики оценки качества распознавания именованных сущностей использовалась метрика f1-score.

Для сравнения предложенного подхода распознавания именованных сущностей с аналогами использовался набор аннотаций научных статей, собранных с arXiv.org и относящихся к тематике компьютерных наук и нефтегазовой отрасли [18]. Набор данных представляет собой 1 000 аннотаций научных публикаций, что в сравнении с публичными наборами данных, которые обычно содержат десятки тысяч текстов, является ограниченной выборкой. Каждая аннотация характеризуется небольшим количеством слов – не более 250, из которых 1–2 слова являются технологическими терминами. Набор данных размечался вручную в соответствие с BIO-разметкой при помощи специализированного инструмента Brat [19]. Далее представлен пример разметки одного предложения.

Предложения, содержащие технологические термины, имеют следующий вид: *«A proper **computing grid-infrastructure** has been constructed at the RDMS institutes for the participation in the running phase of the CMS experiment».* Маска предложения после разметки имеет вид: *«O O **B I** O O O O O O O O O O O O O O O O O O O O».* В представленном случае в предложении присутствует два класса: слова, относящиеся к какой-либо технологии, и остальные слова предложения. Слова *computing* и *grid-infrastructure* принадлежат к искомому классу технологических терминов.

Таблица содержит результаты работы исследуемых методов распознавания именованных сущностей по метрике f1-score при различных размерах обучающей выборки. Значения в таблице представляют собой оценку математического ожидания $\bar{X}$ и оценку стандартного отклонения σ, которые вычисляются по формулам

$$\bar{X} = \frac{x_1 + x_2 + ... + x_n}{n},\ \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}},$$

где x_i – i-е значение метрики f1-score на i-й обучающей выборке, n – количество формирований обучающей выборки, в нашем случае n = 10.

Результаты оценки качества распознавания именованных сущностей по метрике f1-score (оценка среднего отклонения и оценка стандартного отклонения)

Размер обучающей выборки / модели	CRF		Bi-LSTM-CRF		Bi-LSTM-CRF (Scalar Mix)		Предложенный подход	
	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ
200	0,43	0,07	0,54	0,04	0,54	0,05	0,61	0,03
400	0,44	0,04	0,64	0,03	0,66	0,04	0,66	0,04
600	0,47	0,05	0,71	0,02	0,72	0,03	0,77	0,06
800	0,52	0,06	0,74	0,03	0,76	0,05	0,86	0,06

Как видно из таблицы, предложенный метод показывает лучшие результаты по метрике f1-score в задаче по распознаванию именованных сущностей при всех исследуемых размерах обучающей выборки. Единственным исключением является равенство предложенного метода и Bi-LSTM-CRF_SM в ситуации, когда для обучения использовалось 400 текстов.

Данный эффект во многом связан с низким приростом качества распознавания именованных сущностей с помощью метода CRF, а так как он используется как часть предложенного метода, этот эффект накладывается и на него. Данную особенность отчетливо видно на рис. 4: изменения качества распознавания (при переходе от 200 до 400 текстов) для предложенного метода (около 5%) представляют собой что-то среднее между низким ростом у CRF (около 1%) и значительным у методов, основанных на применении Bi-LSTM-CRF (10–12%). Данный эффект показывает, что методу CRF не хватает используемых признаков для качественного распознавания именованных сущностей при объемах выборки 200–400 текстов.

Однако стоит отметить, что на отрезке 400–800 текстов в обучающей выборке прирост качества распознавания для предложенного подхода составил около 20%, в то время как методы, основанные на Bi-LSTM-CRF, показали прирост около 10%. Фактически это свидетельствует о том, что хотя предложенный метод при объемах обучающей выборки до 400 текстов лучше своих аналогов, весь его потенциал раскрывается уже на выборках больше 400 текстов.

На рис. 5 приведена зависимость значения f1-score распознавания именованных сущностей от размера обучающей выборки с учетом разброса предсказаний. Здесь необходимо отметить три наиболее важных момента:

– большой межквартильный размах для метода CRF при размере выборки в 200 текстов, что еще раз подтверждает высказанное ранее предположение о недостатке используемых признаков;

– выброс (отмечено точками) у метода Bi-LSTM-CRF_SM в районе 75% при размере обучающей выборки в 400 текстов, что также косвенно повлияло на равенство результатов работы данного метода и предложенного подхода;

– разнице между 75-м и 25-м квартилzvb относительно медианы у предложенного подхода при обучающей выборкt размером 800 текстов. Расстояние между медианой и 25-м квартилем в несколько раз больше, что является признаком вытянутого влево «хвоста», который занижает значение среднего арифметического (оно равно 86%), тогда так медиана практически равна 90%.

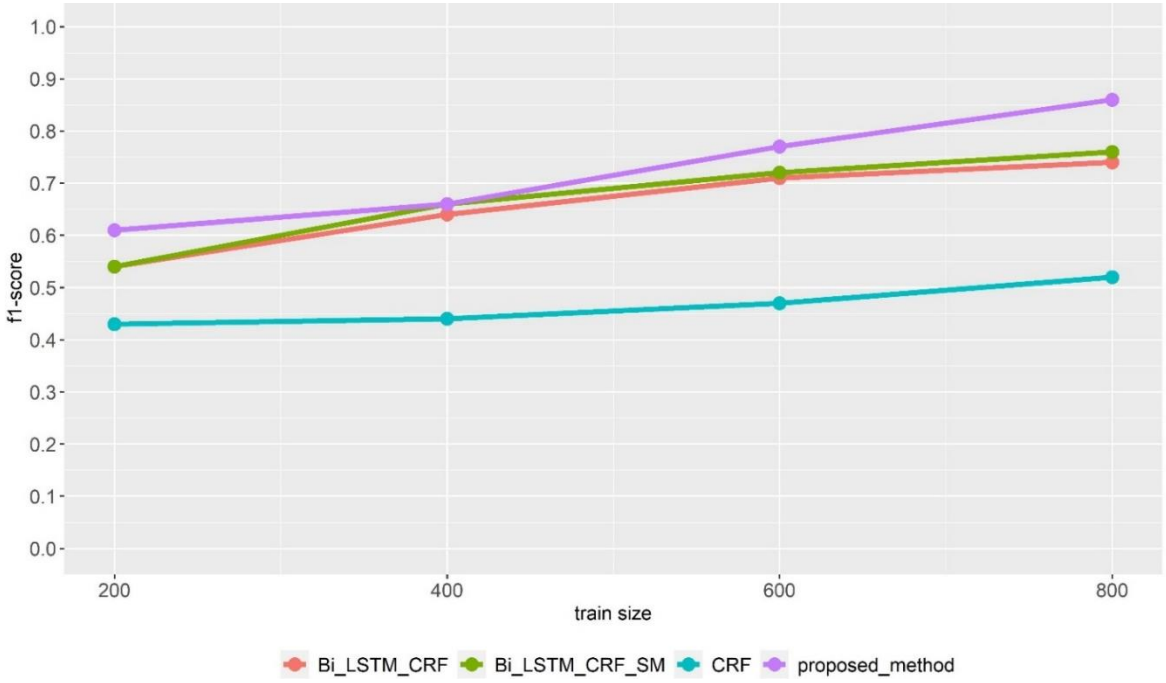

Рис. 4. Зависимость значения f1-score распознавания именованных сущностей от размера обучающей выборки
Fig. 4. Dependence of the f1-score value of named entity recognition on the size of the training sample

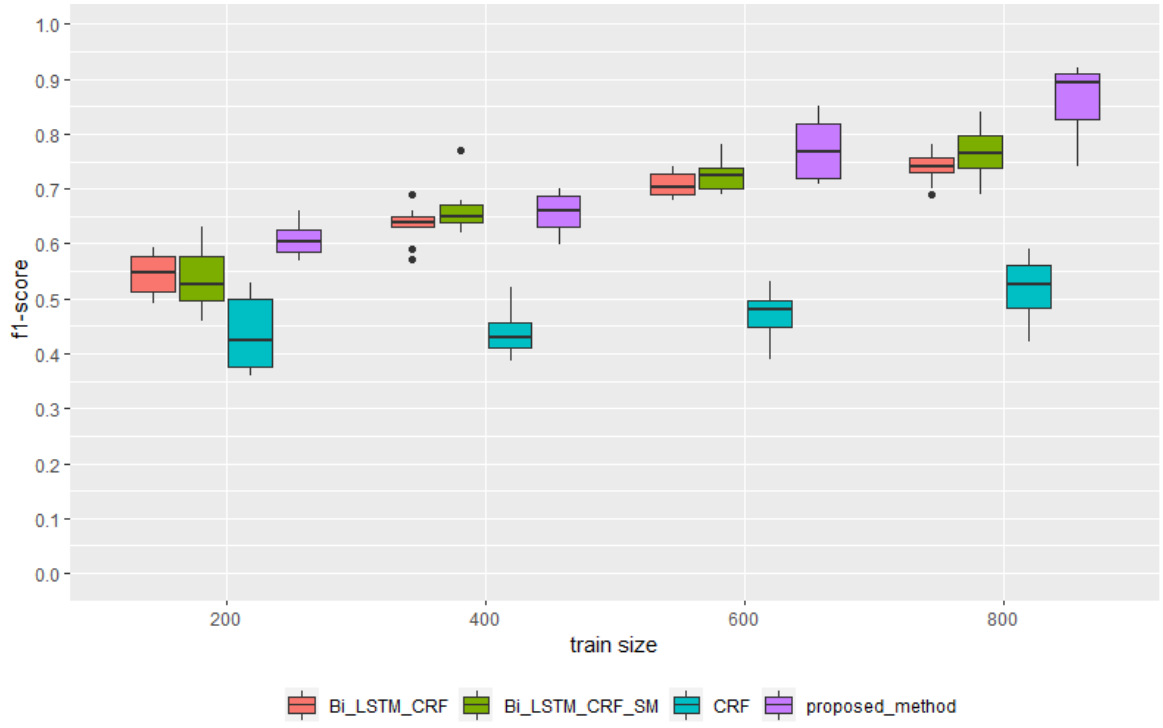

Рис. 5. Зависимость значения f1-score распознавания именованных сущностей от размера обучающей выборки с учетом разброса предсказаний
Fig. 5. Dependence of the f1-score value of named entity recognition on the size of the training sample taking into account the spread of predictions

В целом методы CRF и Bi-LSTM-CRF, используемые по отдельности, не способны достичь точности распознавания именованных сущностей, демонстрируемой предложенным подходом. Благодаря расширению входной информации в виде результатов предсказания CRF для Bi-LSTM-CRF, модель способна лучше обучаться по сравнению с аналогами. В свою очередь, применение Scalar mix, в отличие от комбинирования моделей, не позволяет значительно повысить качество обнаружения именованных сущностей, прирост находится в пределах 1–2%.

Заключение

В рамках исследования продемонстрировано повышение качества распознавания технологических терминов за счет комбинирования статистической и нейросетевой моделей машинного обучения и применения доменно-ориентированной языковой модели для векторного представления научных текстов. Это позволило повысить качество распознавания технологических терминов по метрике f1-score на 12% при обучении на 800 текстах по сравнению с традиционным подходом. Помимо этого, стоит отметить пользу использования техники усреднения значений скрытых слоев языковой модели, что также положительно повлияло на итоговый результат.

В данной работе рассматривались тексты только на английском языке, но имеется возможность масштабирования алгоритма на другие языки, в том числе на русский. Для таких случаев предполагается использование мультиязыковых моделей, обученных на параллельных корпусах данных, благодаря чему одни и те же термины в векторном пространстве находятся достаточно близко друг к другу.

Список источников

1. Nadeau D., Sekine S. A survey of named entity recognition and classification // Lingvisticae Investigationes. 2007. V. 30, № 1. P. 3–26.
2. Marrero M. et al. Named entity recognition: fallacies, challenges and opportunities // Computer Standards & Interfaces. 2013. V. 35, № 5. P. 482–489.
3. Korkontzelos I. et al. Boosting drug named entity recognition using an aggregate classifier // Artificial intelligence in medicine. 2015. V. 65, № 2. P. 145–153.
4. Lafferty J., McCallum A., Pereira F.C.N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001. URL: https://repository.upenn.edu/cgi/viewcontent.cgi?article=1162&context=cis_papers
5. Schuster M., Paliwal K.K. Bidirectional recurrent neural networks // IEEE transactions on Signal Processing. 1997. V. 45, № 11. P. 2673–2681.
6. Jing L., Aixin S., Ray H., Chenliang L. A Survey on Deep Learning for Named Entity Recognition // IEEE Transactions on Knowledge and Data Engineering. 2020. DOI: 10.1109/TKDE.2020.2981314
7. Hossari M., Dev S., Kelleher D.J., TEST: A Terminology Extraction System for Technology Related Terms // ICCAE 2019, February 23–25, 2019. URL: https://arxiv.org/pdf/1812.09541.pdf
8. Jason P.C., Chiu N.E. Named Entity Recognition with Bidirectional LSTM-CNNs // arXiv preprint:1511.08308v5. 2016. URL: https://arxiv.org/pdf/1511.08308.pdf
9. Pennington J., Socher R., Christopher D.M. GloVe: Global Vectors for Word Representation / Computer Science Department, Stanford University. 2014. URL: https://nlp.stanford.edu/pubs/glove.pdf
10. Wang S., Zhou W., Jiang C. A survey of word embeddings based on deep learning // Computing. 2020. V. 102, № 3. P. 717–740.
11. Wang Y. et al. From static to dynamic word representations: a survey // International Journal of Machine Learning and Cybernetics. 2020. V. 11 (4). P. 1–20.
12. Devlin J. et al. BERT: Pre-training of deep bidirectional transformers for language understanding // arXiv preprint arXiv:1810.04805. 2018. URL: https://arxiv.org/pdf/1810.04805.pdf
13. Beltagy I., Lo K., Cohan A. SciBERT: a pretrained language model for scientific text // Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019. DOI: 10.18653/v1/D19-1371
14. Tenney I. et al. What do you learn from context? probing for sentence structure in contextualized word representations // arXiv preprint arXiv:1905.06316. 2019. URL: https://arxiv.org/pdf/1903.10676.pdf
15. Tenney I., Das D., Pavlick E. BERT rediscovers the classical NLP pipeline // Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. P. 4593–4601.
16. Huang Z., Xu W., Yu K. Bidirectional LSTM-CRF models for sequence tagging // arXiv preprint arXiv:1508.01991. 2015. URL: https://arxiv.org/pdf/1508.01991.pdf

17. Vaswani A. et al. Attention is all you need // arXiv preprint arXiv:1706.03762. 2017. URL: https://arxiv.org/pdf/1706.03762.pdf
18. Service for the free distribution of articles in the fields of physics, mathematics, computer science and other. URL: https://arxiv.org/ (accessed: 22.10.2020).
19. Stenetorp P. et al. BRAT: a web-based tool for NLP-assisted text annotation // Proc. of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. 2012. P. 102–107.

References

1. Nadeau, D. & Sekine, S. (2007) A survey of named entity recognition and classification. *Lingvisticae Investigationes.* 30(1). pp. 3–26. DOI: 10.1075/li.30.1.03nad
2. Marrero, M. et al. (2013) Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces*. 35(5). pp. 482–489. DOI: 10.1016/j.csi.2012.09.004
3. Korkontzelos, I. et al. (2015) Boosting drug named entity recognition using an aggregate classifier. *Artificial Intelligence in Medicine*. 65(2). pp. 145–153. DOI: 10.1016/j.artmed.2015.05.007
4. Lafferty, J., McCallum, A. & Pereira, F.C.N. (2001) *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. [Online] Available from: https://repository.upenn.edu/cgi/viewcontent.cgi?article=1162&context=cis_papers
5. Schuster, M. & Paliwal, K.K. (1997) Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*. 45(11). pp. 2673–2681.
6. Jing, L., Aixin, S., Ray, H. & Chenliang, L. (2020) A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*. DOI: 10.1109/TKDE.2020.2981314
7. Hossari, M., Dev, S. & Kelleher, D.J. (2019) TEST: A terminology extraction system for technology related terms. *ICCAE* 2019, February 23–25.
8. Jason, P.C. & Chiu, N.E. (2016) *Named entity recognition with bidirectional LSTM-CNNs*. arXiv preprint:1511.08308v5.
9. Pennington, J., Socher, R. & Christopher, D.M. (2014) GloVe: Global vectors for word representation. *Computer Science Department.* Stanford University, Stanford, CA 94305.
10. Wang, S., Zhou, W. & Jiang, C. (2020) A survey of word embeddings based on deep learning. *Computing*. 102(3). pp. 717–740. DOI: 10.1007/s00607-019-00768-7
11. Wang, Y. et al. (2020) From static to dynamic word representations: a survey. *International Journal of Machine Learning and Cybernetics*. pp. 1–20. DOI: 10.1007/s13042-020-01069-8
12. Devlin, J. et al. (2018) *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805.
13. Beltagy, I., Lo, K. & Cohan, A. (2019) *SciBERT: A pretrained language model for scientific text*. arXiv preprint arXiv:1903.10676.
14. Tenney, I. et al. (2019) *What do you learn from context? probing for sentence structure in contextualized word representations*. arXiv preprint arXiv:1905.06316.
15. Tenney, I., Das D. & Pavlick, E. (2019) *BERT rediscovers the classical NLP pipeli*ne. arXiv preprint arXiv:1905.05950.
16. Huang, Z., Xu, W. & Yu, K. (2015) *Bidirectional LSTM-CRF models for sequence tagging*. arXiv preprint arXiv:1508.01991.
17. Vaswani, A. et al. (2017) *Attention is all you need*. arXiv preprint arXiv:1706.03762.
18. Arxiv.org. (n.d.) *Service for the free distribution of articles in the fields of physics, mathematics, computer science and other*. [Online] Available from: https://arxiv.org/ (Accessed: 22nd October 2020).
19. Stenetorp, P. et al. (2012) BRAT: a web-based tool for NLP-assisted text annotation. *Proc. of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. pp. 102–107.

***Информация об авторах*:**

Кульневич Алексей Дмитриевич – аспирант кафедры теоретических основ информатики Института прикладной математики и компьютерных наук Национального исследовательского Томского государственного университета (Томск, Россия). E-mail: kulnevich94@mail.ru

Кошечкин Александр Алексеевич – аспирант кафедры теоретических основ информатики Института прикладной математики и компьютерных наук Национального исследовательского Томского государственного университета (Томск, Россия). E-mail: kaa1994g@mail.ru

Карев Святослав Васильевич – аспирант кафедры теоретических основ информатики Института прикладной математики и компьютерных наук Национального исследовательского Томского государственного университета (Томск, Россия). E-mail: svyatoslav.karev@live.ru

Замятин Александр Владимирович – профессор, доктор технических наук, заведующий кафедрой теоретических основ информатики, директор Института прикладной математики и компьютерных наук Национального исследовательского Томского государственного университета (Томск, Россия). E-mail: E-mail: avzamyatin@inbox.ru

Вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации. Авторы заявляют об отсутствии конфликта интересов.

***Information about the authors*:**
Kulnevich Alexey Dmitrievich – Post-graduate Student, National Research Tomsk State University (Tomsk, Russian Federation). E-mail: kulnevich94@mail.ru
Koshechkin Alexander Alekseevich – Post-graduate Student, National Research Tomsk State University (Tomsk, Russian Federation). E-mail: kaa1994g@mail.ru
Karev Svyatoslav Vasilyevich – Post-graduate Student, National Research Tomsk State University (Tomsk, Russian Federation). E-mail: svyatoslav.karev@live.ru
Zamyatin Alexander Vladimirovich – Doctor of Technical Sciences, Professor, National Research Tomsk State University (Tomsk, Russian Federation). E-mail: avzamyatin@inbox.ru

Contribution of the authors: the authors contributed equally to this article. The authors declare no conflicts of interests.

Поступила в редакцию 24.07.2021; принята к публикации 28.02.2022

Received 24.07.2021; accepted for publication 28.02.2022