КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2021 Т. 13 № 6 С. 1317–1336
DOI: 10.20537/2076-7633-2021-13-6-1317-1336

Ки&М

МОДЕЛИ ЭКОНОМИЧЕСКИХ И СОЦИАЛЬНЫХ СИСТЕМ

УДК: 004.896, 004.584, 004.91, 519.688

Метод контрастного семплирования для предсказания библиографических ссылок

Ф. В. Краснов[a], И. С. Смазневич, Е. Н. Баскакова

NAUMEN R&D,
Россия, 620028, г. Екатеринбург, ул. Татищева, 49а

E-mail: [a] fkrasnov@naumen.ru

Получено 30.07.2021, после доработки — 14.09.2021.
Принято к публикации 25.09.2021.

В работе рассматривается задача поиска в научной статье фрагментов с недостающими библиографическими ссылками с помощью автоматической бинарной классификации. Для обучения модели предложен метод контрастного семплирования, новшеством которого является рассмотрение контекста ссылки с учетом границ фрагмента, максимально влияющего на вероятность нахождения в нем библиографической ссылки. Обучающая выборка формировалась из автоматически размеченных семплов — фрагментов из трех предложений с метками классов «без ссылки» и «со ссылкой», удовлетворяющих требованию контрастности: семплы разных классов дистанцируются в исходном тексте. Пространство признаков строилось автоматически по статистике встречаемости термов и расширялось за счет конструирования дополнительных признаков — выделенных в тексте сущностей ФИО, чисел, цитат и аббревиатур.

Проведена серия экспериментов на архивах научных журналов «Правоприменение» (273 статьи) и «Журнал инфектологии» (684 статьи). Классификация осуществлялась моделями Nearest Neighbours, RBF SVM, Random Forest, Multilayer Perceptron, с подбором оптимальных гиперпараметров для каждого классификатора.

Эксперименты подтвердили выдвинутую гипотезу. Наиболее высокую точность показал нейросетевой классификатор (95 %), уступающий по скорости линейному, точность которого при контрастном семплировании также оказалась высока (91–94 %). Полученные значения превосходят результаты, опубликованные для задач NER и анализа тональности на данных со сравнимыми характеристиками. Высокая вычислительная эффективность предложенного метода позволяет встраивать его в прикладные системы и обрабатывать документы в онлайн-режиме.

Ключевые слова: контрастное семплирование, анализ цитирования, передискретизация данных, предсказание библиографических ссылок, текстовая классификация, искусственные нейронный сети

COMPUTER RESEARCH AND MODELING
2021 VOL. 13 NO. 6 P. 1317–1336
DOI: 10.20537/2076-7633-2021-13-6-1317-1336

Ки&М

MODELS OF ECONOMIC AND SOCIAL SYSTEMS

UDC: 004.896, 004.584, 004.91, 519.688

Bibliographic link prediction using contrast resampling technique

F. V. Krasnov[a], I. S. Smaznevich, E. N. Baskakova

NAUMEN R&D,
49A, Tatishcheva st., Yekaterinburg, 620028, Russian Federation

E-mail: [a] fkrasnov@naumen.ru

Received 30.07.2021, after completion — 14.09.2021.
Accepted for publication 25.09.2021.

The paper studies the problem of searching for fragments with missing bibliographic links in a scientific article using automatic binary classification. To train the model, we propose a new contrast resampling technique, the innovation of which is the consideration of the context of the link, taking into account the boundaries of the fragment, which mostly affects the probability of presence of a bibliographic links in it. The training set was formed of automatically labeled samples that are fragments of three sentences with class labels «without link» and «with link» that satisfy the requirement of contrast: samples of different classes are distanced in the source text. The feature space was built automatically based on the term occurrence statistics and was expanded by constructing additional features — entities (names, numbers, quotes and abbreviations) recognized in the text.

A series of experiments was carried out on the archives of the scientific journals «Law enforcement review» (273 articles) and «Journal Infectology» (684 articles). The classification was carried out by the models Nearest Neighbors, RBF SVM, Random Forest, Multilayer Perceptron, with the selection of optimal hyperparameters for each classifier.

Experiments have confirmed the hypothesis put forward. The highest accuracy was reached by the neural network classifier (95 %), which is however not as fast as the linear one that showed also high accuracy with contrast resampling (91–94 %). These values are superior to those reported for NER and Sentiment Analysis on comparable data. The high computational efficiency of the proposed method makes it possible to integrate it into applied systems and to process documents online.

Keywords: contrast resampling, citation analysis, data resampling, link prediction, text classification, artificial neural network

Citation: *Computer Research and Modeling*, 2021, vol. 13, no. 6, pp. 1317–1336 (Russian).

Введение

Современные научные исследования невозможны без соотнесения полученных результатов с работами других ученых. Корректность цитирования важна для всех участников научного сообщества — авторов статей, редакций журналов, специалистов по наукометрии, научных администраторов. В работе изучается возможность формирования автоматической рекомендации для указания необходимости библиографической ссылки в определенном фрагменте текста при создании научной статьи.

Цель исследования — предложить алгоритм поиска в научной статье фрагментов с недостающими библиографическими ссылками и проверить гипотезу: контрастное семплирование текста при добавлении специальных признаков позволяет строить высокоточный бинарный классификатор для определения вероятности недостающих библиографических ссылок в тексте научных статей. Постановка задачи развивает такие хорошо исследованные направления анализа текста, как выявление именованных сущностей и анализ тональности.

Существующие исследования по анализу цитирования затрагивают смежные задачи: способы анализа связей, формы представления результатов, определение важных фрагментов упомянутых статей. Ряд исследований, нацеленных на улучшение качества цитирования, посвящены поиску наиболее релевантной статьи. Однако работ по рекомендации добавления ссылок и выбора позиции в тексте до настоящего времени не проводилось.

Связи научных статьей с результатами других ученых выражаются в их упоминании в виде библиографической ссылки. Обоснование необходимости установления таких связей между различными исследованиями изучается специалистами по наукометрии, которые формулируют различные теории цитирования. В частности, согласно нормативной теории, «ссылки в научных работах делаются для того, чтобы обозначить работы, являющиеся основой для излагаемого исследования, описывающие используемые методы исследования, связанные тематически и необходимые для обсуждения полученных результатов» [Акоев и др., 2014]. Эта теория опирается на принципы научной этики, сформулированные в [Merton, 1973]: общедоступность научных достижений после их публикации, независимость научного значения результатов от личности автора (в том числе национальности, положения, характера), бескорыстность научной деятельности, критичность к собственным и чужим результатам. С точки зрения рефлексивной теории цитирования ссылки между научными работами являются индикаторами состояния науки, они отражают ее характер и позволяют создавать ее формализованное представление, например, строить карты науки.

Таким образом, в корректности научного цитирования заинтересовано всё научное сообщество, как исследователи, создающие статьи о своих результатах, так и администраторы, отслеживающие положение дел в различных отраслях научных знаний. Упоминание достижений других ученых является одним из базовых требований при построении научных текстов и обязательным критерием качества статей с точки зрения редакций научных журналов. Эти требования фиксируются в методических руководствах по составлению академических текстов [Emerson, Rees, MacKay, 2005; Gray et al., 2008; Pears, Shields, 2019] и подтверждаются на практике, например результатами исследований публикационной активности в высокорейтинговых международных журналах [Arsyad et al., 2020].

При добавлении цитат в научную статью выбор источников и мест в тексте, подходящих для указания ссылок, делает сам автор научной работы, и этот процесс в настоящее время не автоматизирован. Поэтому, несмотря на имеющиеся широкие возможности поиска информации, составление обзора текущего состояния рассматриваемой отрасли науки, в котором требуется указание многочисленных исследований по предмету, является достаточно трудоемким этапом создания научной статьи. В настоящей работе исследуется возможность создания рекомендательного алгоритма, позволяющего обнаружить недостающие в тексте библиографические ссылки,

то есть выявлять те фрагменты текста научной статьи, где необходимо упоминание другой исследовательской работы. Для этого решается задача оценки вероятности необходимости ссылки в различных фрагментах текста научной статьи с использованием подхода машинного обучения с привлечением учителя, при котором в качестве обучающих данных рассматриваются массив текстов научных статей с указанными в них библиографическими ссылками.

Обзор

Анализ ссылок в массиве документов проводится в рамках исследований по различным областям знаний. Задача анализа ссылок может быть рассмотрена как междисциплинарная. Типологическая характеристика ссылок интересует ученых-филологов. В работе [Носовец, 2011] исследуются гиперссылки в интернет-СМИ и подразделяются в зависимости от их целевого назначения: контент-ссылки, уточняющие ссылки, сервисные, информационные, коммуникационные, рекламные, а также отмечаются внешние различия в оформлении ссылок разного типа. В статье [Стройков, 2010] вводится понятие электронного лексикографического гипертекста и делается классификация ссылок: однонаправленные и двунаправленные (перекрестные), внутренние (внутритекстовые) и внешние (межтекстовые). В библиотечном деле на основе ссылочности формируются различные библиометрические индикаторы.

В сфере наукометрии ссылки являются основополагающим понятием: они позволяют выявлять и анализировать научные сообщества [Chandrasekharan et al., 2021], вычислять степень влиятельности отдельных ученых и научных коллективов [Yang, Wang, 2015], выявлять перспективные направления исследований [Trujillo, Long, 2018] и вычислять импакт-фактор журналов, а также оценивать сами статьи [Herrmannova, Knoth, 2016].

Специалисты по наукометрии делают различные выводы на основе исследования ссылок между документами. Например, в работе [Bornmann, Wray, Haunschild, 2020] с помощью анализа цитирования прослеживается распространение влияния упоминаемых концепций предметной области; в другой статье этого же автора [Bornmann, Wagner, Leydesdorff, 2018] предлагается метод оценки вклада стран в мировую науку на основе анализа ссылок из наиболее авторитетных работ. Алгоритмы Tree of Science и его модификация SAP [Valencia-Hernández et al., 2020] позволяют проследить развитие научных отраслей, представляя результаты в виде дерева, листьями которого выступают самые новые работы, а корнями — основополагающие. В статье [Miura, Asatani, Sakata, 2020] исследуется феномен «отложенного узнавания» научных работ, когда ценность новой концепции, которая оказывается прорывной для решения некоторой научной задачи, осознается научным сообществом не сразу; анализ цитируемости помогает выявить такие пары работ: автора прорывной концепции («спящая красавица») и автора ее первого упоминания («принца»). Анализ цитирования широко применяется и в области патентного анализа, в том числе с целью обнаружения инновационных решений и прорывных технологий [Lai et al., 2021].

Различные методы анализа цитирования дают разные результаты с точки зрения качества получаемой таксономии сферы научных исследований [Klavans, Boyack, 2017]. В социологии, как и в наукометрии, ссылочная связь между научными работами и между их авторами рассматривается в основном как связь через атрибуты, поскольку источником данных являются библиографические списки, прилагаемые к статьям. При этом подсчет числа цитирований с целью определения значимости научной работы без учета того, в каком контексте была упомянута та или иная работа, в настоящее время справедливо подвергается критике [Акоев и др., 2014; MacRoberts M., MacRoberts B., 2018]. Для того чтобы оценка важности статьи на основе анализа цитирования была корректной, можно либо модифицировать алгоритмы ранжирования статей, принимая во внимание оценки, относящиеся к авторам и к журналам, либо учитывать информацию, содержащуюся в контексте ссылки на научную работу.

В то же время само по себе рассмотрение контекста ссылок при анализе цитирования приводит к различным выводам. В обзорной статье [Tahamtan, Bornmann, 2019] обсуждаются различные возможности такого анализа, в том числе предлагаются критерии для классификации ссылок. В работе [Manggala et al., 2021] исследуются типы использования ссылок, и этой информацией обогащается граф цитирования. В данном случае источником данных выступают интервью, однако подобные задачи решаются и методами автоматического анализа текста с использованием различных алгоритмов обработки естественного языка (Natural Language Processing, NLP). Например, в ряде работ исследуется возможность определения нужного фрагмента в тексте упоминаемой статьи, то есть того, который имелся в виду при цитировании [Huang, Krylova, 2020; Chandrasekaran et al., 2019]. А в других работах [Medić, Snajder, 2020] решается задача выбора научной статьи, наиболее подходящей для цитирования в данном фрагменте текста, учитывая как локальный контекст, так и контекст документа.

В рамках данной работы задача автоматического определения необходимости ссылки рассматривается в следующей постановке: требуется обнаружить в тексте научной статьи те фрагменты (предложения), где ссылка отсутствует, но необходима, на основе данных обучающей выборки — наборов фрагментов со ссылками и без.

Задача классификации текстовых фрагментов с точки зрения контекста содержащихся в них ссылок методологически сходна с задачей анализа тональности (Sentiment Analysis), при котором автоматически определяется эмоциональная окраска текстов, в основном коротких сообщений в социальных сетях и новостных сообщений в массмедиа. Кроме деления на позитивные и негативные фрагменты, подход анализа тональности используется и для выделения других классов. Например, в работе [Aljuaid et al., 2021] подход анализа тональности используется для классификации ссылок на важные и неважные. В другом исследовании [Prester et al., 2021], также посвященном проблеме качества научных статей и решению проблем цитирования, похожий подход используется для анализа научных статей с точки зрения значимости предложенных в них идей.

Другим близким направлением исследований является выделение из текста именованных сущностей (Named Entity Recognition, NER) через предсказывание классификаторами. Похожая задача решается авторами [Fu, Huang, Liu, 2021], NER рассматривается в подходе Span Prediction. В работе [Wang et al., 2021] для улучшения точности учитывается внешний контекст предложения, для чего находятся семантически близкие тексты через поисковый запрос по рассматриваемому предложению. В [Ziyadi et al., 2020] та же задача рассматривается при малом числе доступных образцов и решается в два этапа, первым из которых является обучение для определения предложений, содержащих сущности, без определения границ этих сущностей. Автор [Li, 2021] проводит аналогию между обнаружением объектов в компьютерном зрении и NER в текстах, предлагая схожий двухэтапный метод, на первом шаге которого для определения областей с наиболее вероятным содержанием именованных сущностей областям дается соответствующая оценка (entityness).

При решении исследуемой задачи методами машинного обучения важную роль играет конструирование набора признаков (feature engineering). В работе [Mozharova, Loukachevitch, 2016] задача NER решается в два этапа, и набор признаков для второго из них формируется совокупно на основе статистики документа, глобальной статистики по коллекции, а также данных первого этапа. В уже упомянутой работе [Prester et al., 2021] для поиска «значимой идеи», помимо эмбеддингов слов, учитываются дополнительные признаки на основе синтаксических, семантических и контекстуальных характеристик цитирующих предложений. Нейросетевые классификаторы выделяют значимые признаки автоматически [Dernoncourt, Lee, Szolovits, 2017], однако добавление новой информации о входных данных по-прежнему способно улучшить результат.

Гипотеза

В рамках данного исследования была проверена следующая гипотеза: контрастное семплирование текста с учетом добавления специальных признаков позволяет строить высокоточный бинарный классификатор для обнаружения (вероятностного определения) недостающих ссылок в тексте научных статей.

Базовый вариант алгоритма

- Обучающая выборка состоит из размеченных семплов (фрагментов со ссылкой или без).
- Позитивные семплы имеют вид «ссылка и ее контекст», негативные — фрагмент текста без ссылок.
- Контекст ссылки ограничивается содержащим ее предложением.
- Пространство признаков формируется автоматически на основе словарной статистики в рамках модели BoW (Bag-of-Words, «мешок слов»).

Модификации, повышающие эффективность алгоритма

- Контекст ссылки не ограничивается одним предложением и расширяется до фрагмента большего объема (три предложения).
- При формировании семплов из текста используется скользящее окно длиной в три предложения с шагом в одно предложение.
- Требование изолированности (контрастности) негативных семплов: в исходном тексте негативные семплы находятся достаточно далеко от позитивных. Алгоритм:
  - рассматриваются фрагменты, размер которых превышает размер семпла;
  - отбираются такие фрагменты без ссылок, к которым в исходном тексте не примыкают предложения со ссылками;
  - для семплов выбирается центральная часть отобранных фрагментов.
- Словарь включает знаки препинания, ограничения по частотности слов отсутствуют.
- Учитываются дополнительные признаки — именованные сущности: числа, цитаты, ФИО.

Методика

Задача определения недостающей ссылки формализуется следующим образом: обнаружить фрагменты текста, где ссылки не хватает либо она лишняя. Текст делится на фрагменты, и далее происходит бинарная классификация относительно критерия необходимости подкрепления изложенных фактов или сформулированных утверждений. Такое деление текста похоже на классификацию в задаче определения полярности при анализе тональности текста, когда требуется определить, является ли каждый фрагмент текста эмоционально-окрашенным отрицательно или положительно.

В то же время исследуемая задача близка к задаче NER — в том смысле, что в тексте необходимо обнаружить фрагменты, которые наиболее вероятно содержат определенную сущность — ссылку на научную работу. Для обнаружения недостающих ссылок при обучении модели может быть использован контекст уже известных ссылок, так же, как контекст используется в задаче

NER. В то же время в обучающей выборке фрагменты явно делятся на «позитивные» и «негативные» по отношению к наличию ссылки, и в этом решение методически сходно с определением полярности эмоциональной окраски. Таким образом, предлагаемый метод частично комбинирует подходы NER и Sentiment Analysis, однако существенно отличается от каждого из них.

Отличительной особенностью предлагаемого метода является явное указание на фрагменты с отрицательной необходимостью ссылки — аналогичный подход известен как Negative Sampling и применяется при построении векторных моделей на основе дистрибутивной семантики. Однако в данном случае вводится дополнительное требование, усиливающее выразительность негативного семпла, состоящее в отсутствии непосредственного примыкания позитивного семпла к негативному, то есть требование удаленности друг от друга семплов с разной полярностью в исходном тексте.

Локальные и глобальные контексты учитываются современными нейросетевыми архитектурами анализа текста. Так как текст, по сути, является однонаправленным списком термов, то под контекстами, согласно [Gallant, 1991; Huang et al., 2012], принято понимать окрестность до или после рассматриваемого терма. Модель дистрибутивной семантики word2vec [Mikolov et al., 2013] использует окно в 5–7 термов для локального контекста, GloVe [Pennington, Socher, Manning, 2014] использует глобальный контекст документа, a fasttext [Bojanowski et al., 2017; Joulin et al., 2016] — субсловарные n-грамы. В сверточных сетях увеличение размера контекста приводит к значительному росту размерности тензоров и, как следствие, параметров модели, что требует увеличения размеров коллекций. В моделях архитектуры Transformer контекст учитывается с помощью «механизма внимания» (attention), и локальный контекст смешивается с более широким контекстом (BERT [Devlin et al., 2018], GPT-3 [Brown et al., 2020], ruGPT-3 [Russian GPT-3]).

Важно отдельно отметить, что все вышеперечисленные алгоритмы не учитывают естественных структурных единиц текстов — предложений и параграфов, так как настраиваются на размер контекста, фиксированный числом слов, а размер предложений и параграфов может меняться.

Для решения исследуемой задачи обнаружения недостающих ссылок были задействованы как методы машинного обучения, относящиеся к информатике и математике, так и знания из области лингвистики о принципах построения текста. Наличие библиографической ссылки в тексте научной статьи означает, что автору в данном контексте требуется сослаться на другую работу: это может быть упоминание с целью цитирования, придания веса аргументам, ознакомление читателя с тематикой проблемы и т. д. Размер контекста, обосновывающий включение библиографической ссылки, определяется только автором, то есть человеком. Таким образом, необходимо сформулировать оптимальную стратегию формирования контекстов, позволяющую автоматически учитывать неопределенность размера фрагмента текста, ограничивающего окрестность библиографической ссылки.

Наилучшим вариантом представляется ситуация, когда границы контекста каждой ссылки совпадают с границами законченной мысли, к которой относится данная ссылка. При этом смысловой единицей текста может быть как одно, так и несколько предложений, что затрудняет выбор длины контекста. Тем не менее для приближения к указанной цели в предлагаемом алгоритме контекстом рассматриваемой сущности (ссылки) считался фрагмент, формирующийся на основе естественных структурных единиц теста: при разбиении текста на семплы использовалось скользящее окно, размер которого определяется числом предложений, а не слов, как это делается в нейросетевых алгоритмах.

В представленном алгоритме не учитывается последовательность термов в контексте и каждый из фрагментов рассматривается в модели «мешка слов». Однако семплирование с по-

мощью скользящего окна позволяет учитывать направление контекста и последовательность предложений внутри контекста.

Задача обнаружения недостающих ссылок решалась в данной работе методами автоматической классификации: для каждого фрагмента научной статьи определялась вероятность наличия в нем библиографической ссылки. Для этого необходимо было найти такой классификатор, который с заданной точностью относит фрагменты текста к классу Y (классу фрагментов со ссылками).

Формальная постановка задачи

Дана коллекция текстовых документов $D = \{d_0, \ldots, d_N\}$ такая, что каждый документ состоит из фрагментов $d_i = \{s^{i_0}, \ldots, s^{i_M}\}$. Фрагменты могут накладываться один на другой и быть разного размера, но каждый фрагмент принадлежит одному классу $Y = \{y_0, \ldots, y_K\}$, с метками класса $y_j \in \{0, 1\}$. Фрагмент s^{i_j} состоит из последовательности слов (термов) разной длины. Метка класса y_j соответствует тому, содержит или нет данный фрагмент s^{i_j} библиографическую ссылку. Задача данного исследования состоит в том, чтобы найти стратегию формирования фрагментов s^{i_j}, функцию классификатора и оптимальный набор гиперпараметров, дающих наибольшую точность определения меток класса y_j.

Решение задачи в указанной формальной постановке проводилось на основе базового алгоритма, который постепенно модифицировался в рамках методики последовательного улучшения точности. Базовый алгоритм состоял из этапов: предобработка текста, разметка текста, классификация. Для увеличения точности были добавлены этапы обработки именованных сущностей и формирования семплов. Общая схема базового и улучшенного алгоритмов изображена на рис. 1.

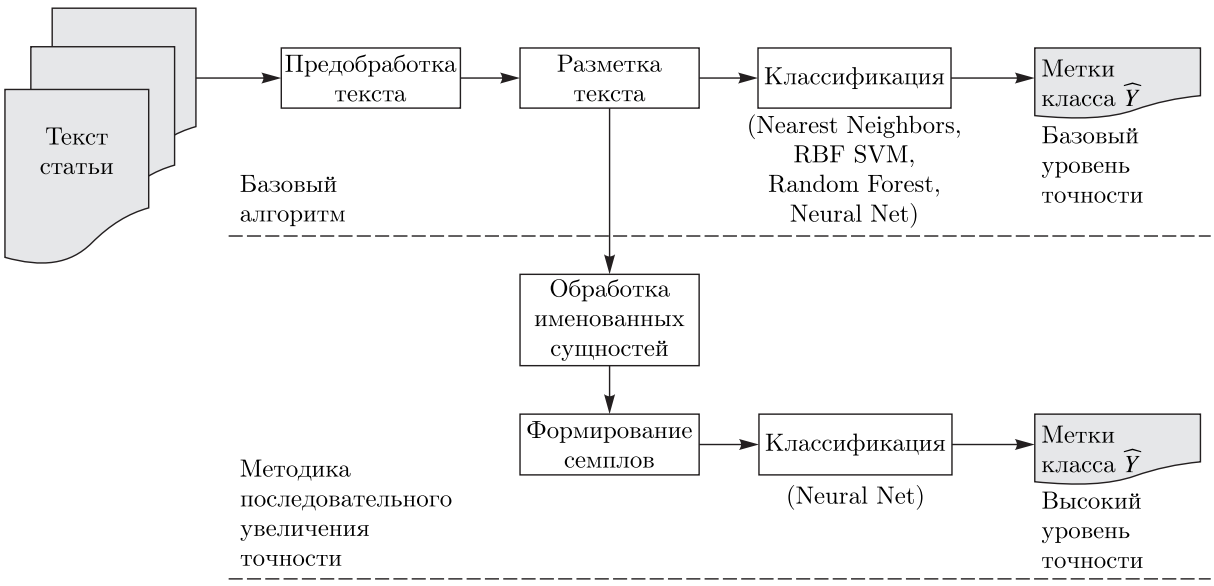

Рис. 1. Общая схема базового и полного алгоритмов классификации фрагментов текста

Этапы полного (улучшенного) алгоритма состояли из следующих операций.

1. Предобработка текстов:

- очистка текста: удаление служебных символов (табуляция, перевод строки и т. д.); удаление служебных слов (название журнала, ISBN и т. д.), удаление списка литературы;
- токенизация: сегментация текста на предложения; приведение термов к начальной форме. Предложение представляет собой массив токенов, включающий слова, знаки препинания и цифры.

2. Разметка текста:

   - для каждой статьи добавляются метки начала и окончания статьи;
   - для каждого предложения добавляется метка принадлежности к классу «со ссылкой», если в предложении есть библиографическая ссылка в квадратных скобках; если библиографическая ссылка отсутствует, то предложение должно быть добавлено к классу «без ссылки». После разметки предложений библиографические ссылки в квадратных скобках должны быть удалены из текста.

3. Обработка именованных сущностей (NER):

   - выделение из текста именованных сущностей: ФИО, числа, даты, цитаты и аббревиатуры;
   - замена ФИО, чисел и дат на специальные метки; добавление специальных меток к цитатам и аббревиатурам.

4. Формирование семплов:

   - объединение предложений в семплы с помощью скользящего окна, размер и шаг которого измеряются в предложениях;
   - разметка семплов по правилам: если хотя бы в одном предложении семпла есть метка «со ссылкой», то необходимо считать этот семпл позитивным; если все предложения семпла, а также четыре соседних предложения (по два с каждой стороны) имеют метку «без ссылки», то считать семпл негативным.

5. Классификация: обработка подготовленных наборов семплов с помощью различных классификаторов с подбором для каждого из них оптимальных гиперпараметров.

Эксперимент

Для проверки гипотез были выбраны два архива научных журналов «Правоприменение» (273 статьи 2017–2019 гг.) и «Журнал инфектологии» (684 статьи 2009–2019 гг.). Файлы статей были переведены из PDF в текстовый формат с помощью библиотеки Apache Tika [Apache Tika].

Базовый алгоритм

Тексты статей были разделены на предложения, для чего использовалась библиотека Natasha [Проект Natasha]. Предложения были взяты только из содержательной части статей, из рассмотрения исключались списки литературы, аннотации и прочие вспомогательные разделы. В результате было получено два набора предложений: набор Коллекция1 из 13 724 предложений и набор Коллекция2 из 86 802 предложений.

Далее производилась разметка данных. Из коллекции Коллекция1 было выделено 1473 предложений, содержащих библиографические ссылки (соответствующие спискам литературы), и 12 251 предложений, не содержащих таких ссылок. Аналогично из второй коллекции

было выделено 6477 предложений со ссылками и 80 325 предложений без ссылок. Таким образом, было получено два размеченных набора данных для дальнейшего эксперимента.

Стоит отметить, что полученные наборы данных Коллекция1 и Коллекция2 оказались не сбалансированными по числу элементов с разными метками классов. Для балансировки данных могут применяться разные стратегии, однако поскольку не всем классификаторам требуется сбалансированность обучающей выборки (этого требуют только линейные классификаторы), то балансировка коллекций на данном этапе не производилась.

Для обеих коллекций был выполнен частотный анализ состава данных. На рис. 2 представлены результаты сравнения распределений плотности термов в коллекции и в предложениях для различных меток класса («со ссылкой» и «без ссылки»). Под плотностью понимаются доля предложений различной длины в общем числе предложений с заданной меткой (верхний ряд гистограмм) и количество термов с разной частотой встречаемости в подмножестве коллекции с заданной меткой (нижний ряд). Из рисунка видно, что распределения меток классов не содержат явных аномалий.

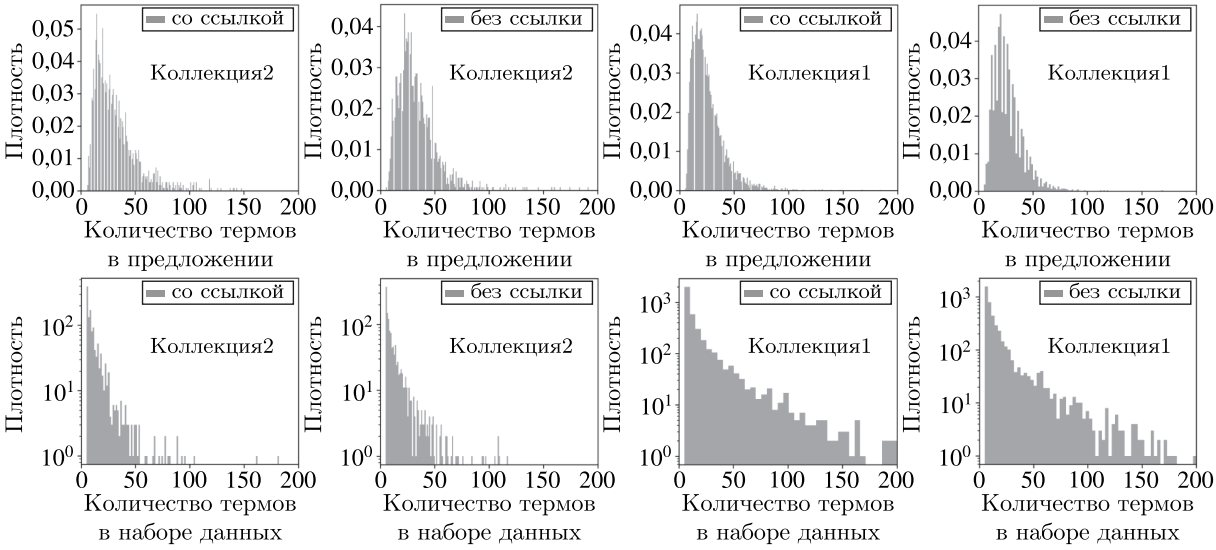

Рис. 2. Частотные характеристики коллекций. Под плотностью понимаются: в верхнем ряду гистограмм — доля предложений различной длины в общем числе предложений с заданной меткой; в нижнем ряду — количество термов с разной частотой встречаемости в подмножестве коллекции с заданной меткой

Для создания векторной модели предложений была построена матрица «терм-документ» со значениями функции TF-IDF. Размерность пространства признаков составила 2493 для набора Коллекция1 и 7261 для набора Коллекция2.

Для решения задачи классификации были выбрано несколько моделей-кандидатов, а именно: Nearest Neighbours, RBF SVM, Random Forest, Fully Connected Neural Net (использовалась реализация методов в библиотеке Scikit-learn [Scikit-learn]). Такой выбор моделей должен был охватить различные нелинейные функции для построения гиперплоскости границы между метками класса.

Для каждого из рассматриваемых классификаторов проводился подбор оптимальных гиперпараметров с использованием метода поиска по сетке с полным перебором.

Результат работы каждого из классификаторов-кандидатов, выраженный в значениях мер точности F1, Precision и Recall, показан в таблице 1. На данном этапе эксперимента для обеих коллекций наилучший результат показал нейросетевой классификатор.

Таблица 1. Мера точности F1 в зависимости от модели для классификации предложений

Набор данных	Классификатор	F1	Precision	Recall
Коллекция2	Nearest Neighbors (Uniform)	0,55	0,71	0,5
	RBF SVM (C = 0,5)	0,67	0,5	1
	Random Forest (300)	0,69	0,7	0,67
	Neural Net ($FC100 \times 5$)	**0,7**	**0,68**	**0,81**
Коллекция1	Nearest Neighbors (Uniform)	0,33	0,6	0,23
	RBF SVM (C = 0,5)	0,7	0,49	1
	Random Forest (300)	0,53	0,60	0,48
	Neural Net ($FC100 \times 5$)	**0,7**	**0,49**	**1**

Далее полученные значения меры точности F1, приведенные в таблице 1, рассматривались как базовые для сравнения с результатами при улучшении алгоритма.

Улучшенный алгоритм

На рис. 3 показана схема улучшенного алгоритма. К базовому алгоритму был добавлены этапы формирования семплов и обработки именованных сущностей.

Для формирования набора семплов в соответствии с предложенной методикой был создан параметризованный конвейер. Под семплами понимаются размеченные текстовые фрагменты из нескольких предложений, определяющих контекст ссылок (позитивные семплы) либо представляющих собой образцы текста без ссылки (негативные семплы). К семплам разных классов предъявляется требование контрастности, состоящее в удаленности позитивных и негативных семплов друг от друга в исходном тексте. Для формирования семплов использовалось скользящее окно размером в 3 предложения с шагом в 1 предложение. Для выполнения требования контрастности проводилась проверка: для каждого фрагмента без ссылки (потенциального негативного семпла) задавалась окрестность, в которую не должны входить позитивные семплы. В ходе эксперимента было установлено, что наибольшая точность достигается при размере семплов, равном 3 предложениям, и окрестности негативных семплов размером по 2 предложения слева и справа.

Алгоритм формирования семплов представлен на рис. 4.

Другим улучшением алгоритма, повлиявшим на точность классификации, было выделение из текста именованных сущностей: имена персон (фамилии и инициалы), числа, даты, цитаты и термины в кавычках, аббревиатуры. Гипотеза об улучшении точности классификации при учете этих сущностей базировалась на экспертном знании предметной области: библиографическая ссылка в тексте может соседствовать рядом с цитатой (в этом случае ссылаются на источник), с вводимым впервые термином либо новой аббревиатурой (что также может сопровождаться ссылкой на работу автора термина либо аббревиатуры); ссылка может сопровождать включенное в повествование упоминание автора некоторого исследования, числа (если приводятся факты, требующие подкрепления), даты (при упоминании научного события).

Указанные именованные сущности обнаруживались в тексте с помощью регулярных выражений. Затем для имен, чисел и дат делалась замена найденных сущностей на метки «ФИО», «ЦЦЦЦ» и «ДДДД» соответственно, а в случае терминов, цитат и аббревиатур метки добавлялись рядом с ними («КККК» и «АБВ» соответственно). Такое разделение было сделано потому, что в цитатах, терминах и аббревиатурах содержится семантическая информация, которая не должна быть утрачена при формировании векторов признаков для семплов.

В соответствии с результатами, представленными в таблице 1, наибольшую точность на предыдущем этапе эксперимента показал классификатор, основанный на нейронной сети, поэтому на данном этапе в качестве классификатора для вычисления вероятности ссылок был

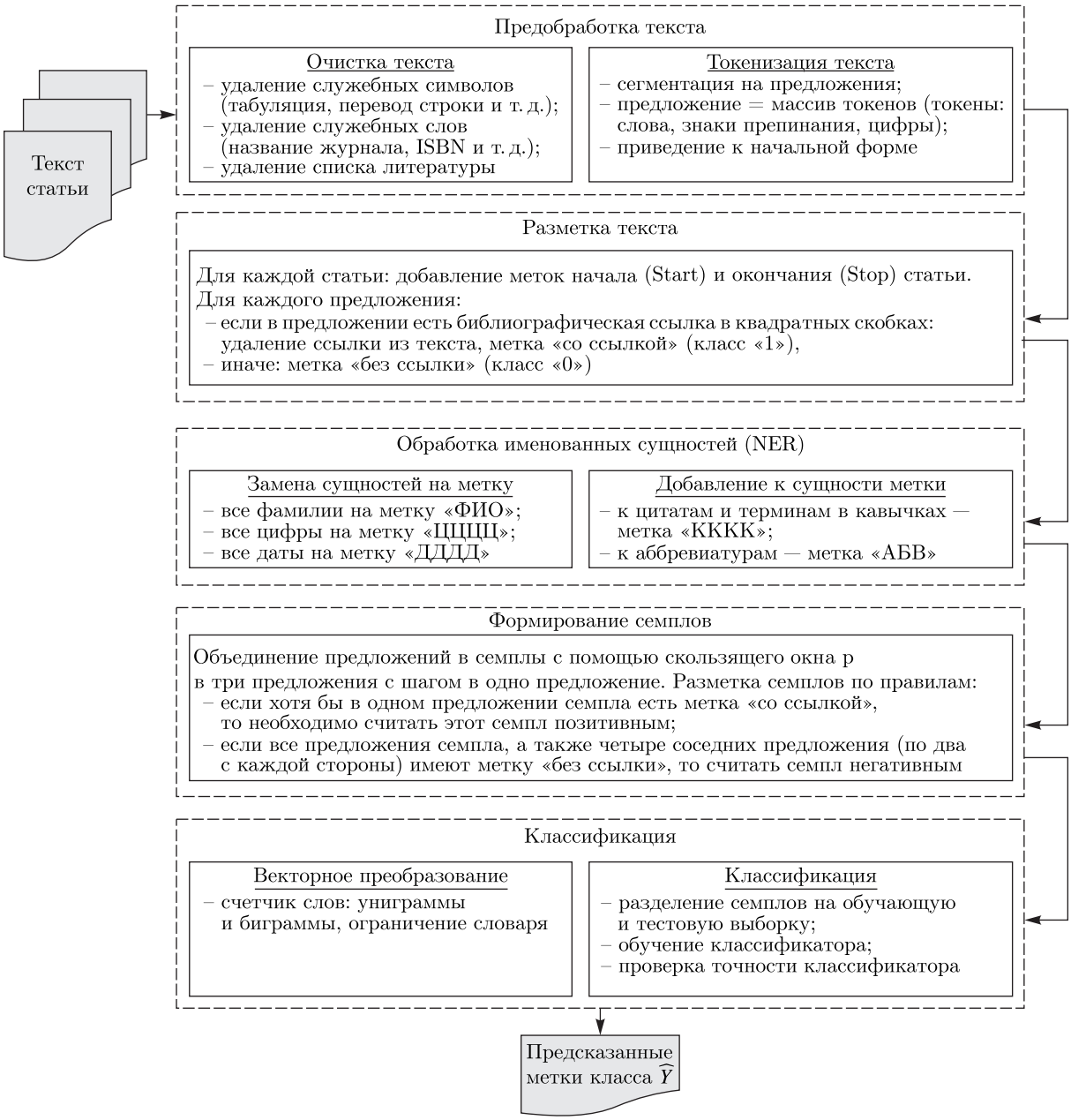

Рис. 3. Улучшенный процесс классификации семплов в рамках методики последовательного увеличения точности оценки вероятности

использован многослойный персептрон (метод MLPClassifier библиотеки Scikit-learn [Scikit-learn]).

Схема классификатора представлена на рис. 5. На вход сети подается тренировочная выборка, размер словаря которой ограничен 30 000 токенами. Размер первого скрытого слоя — 256 нейронов, второго — 8 нейронов. Выходной слой состоит из 1 нейрона: если для обработанного семпла вероятность ссылки больше либо равна 0,5, то принимается решение о принадлежности семпла к позитивному классу, иначе — к негативному. В качестве функции активации используется функция Relu; подбор весов осуществляется с помощью оптимизированного ал-

Алгоритм 1. Процесс формирования семплов

Вход: $Sents$ — обработанные предложения, из которых удалены ссылки;
$y\{0, 1\}$ — метки классов;
$Start$ — метки начала каждой статьи;
$Stop$ — метки окончания каждой статьи.

Выход: $SentsPos$ — набор позитивных семплов;
$SentsNeg$ — набор негативных семплов.

```
1  i := 0;
2  для всех y in(3, M)
3      отобрать последовательно фрагмент из 3 предложений Sents[i], Sents[i + 1], Sents[i + 2]:
4          если хотя бы в одном предложении есть ссылка: сумма (y[i], y[i + 1], y[i + 2]) >= 1
           и предложение не имеет меток Start и Stop;
5              включить фрагмент в набор SentsPos;
6          если ни в одном предложении нет ссылки: y[i] == 0, y[i + 1] == 0, y[i + 2] == 0
           и предложение не имеет меток Start и Stop, то проверить наличие ссылок
           в 2 предложениях слева и справа от фрагмента:
7              если во всех предложениях окрестности нет ссылок: y[i − 1] == 0, y[i − 2] == 0,
               y[i + 3] == 0, y[i + 4] == 0:
8                  включить фрагмент в набор SentsNeg, иначе пропустить.
```

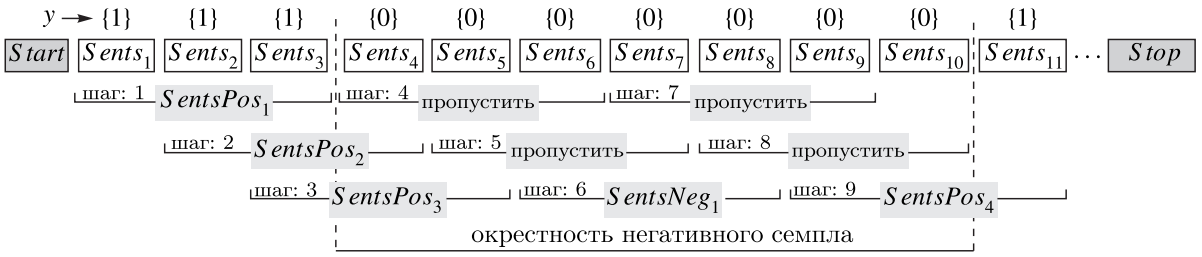

Рис. 4. Алгоритм формирования контрастных семплов для контекста ссылки размером в 3 предложения и окрестности негативного семпла размером в 7 предложений

горитма стохастического градиентного спуска Adam (Adaptive moment estimation, [Kingma, Ba, 2014]).

В таблице 2 представлены результаты работы данного классификатора на обучающей и тестовой выборках в экспериментах на обеих коллекциях.

Таблица 2. Точность классификатора MLPClassifier (256×8) на обучающей и тестовой выборках из двух коллекций

Коллекция	Выборка	F1	Precision	Recall
Коллекция1	Обучающая выборка	1	1	1
	Тестовая выборка	0,94	0,95	0,93
Коллекция2	Обучающая выборка	1	1	1
	Тестовая выборка	0,95	0,93	0,91

На рис. 6 показана зависимость точности нейросетевого классификатора MLPClassifier, выраженной функцией F1-score weighted, от номера эпохи обучения.

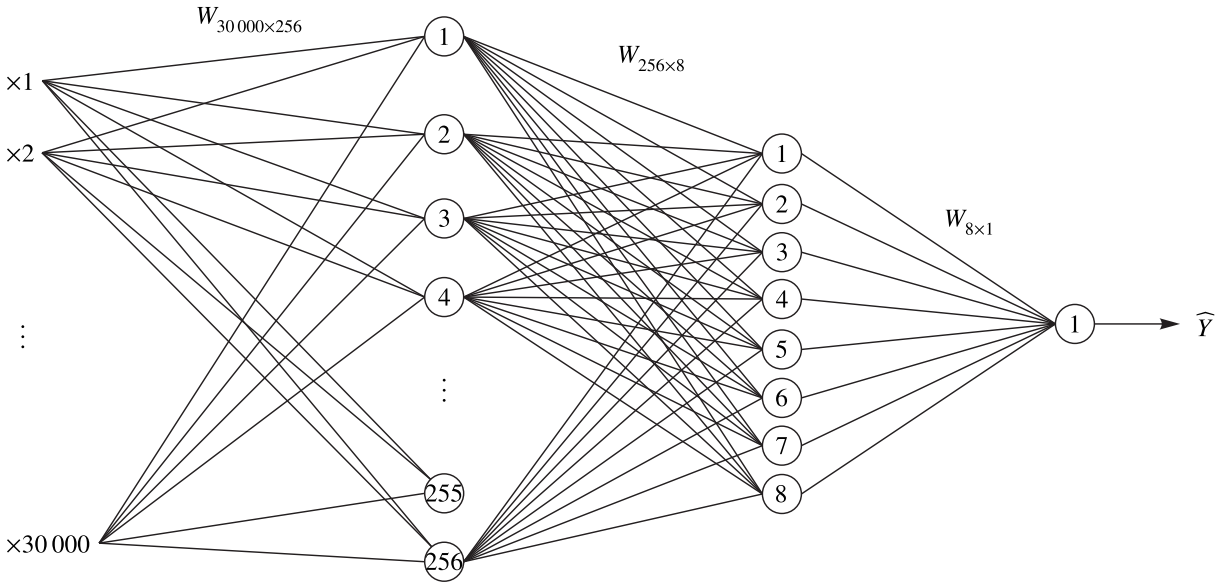

Рис. 5. Схема классификатора MLPClassifier (256×8)

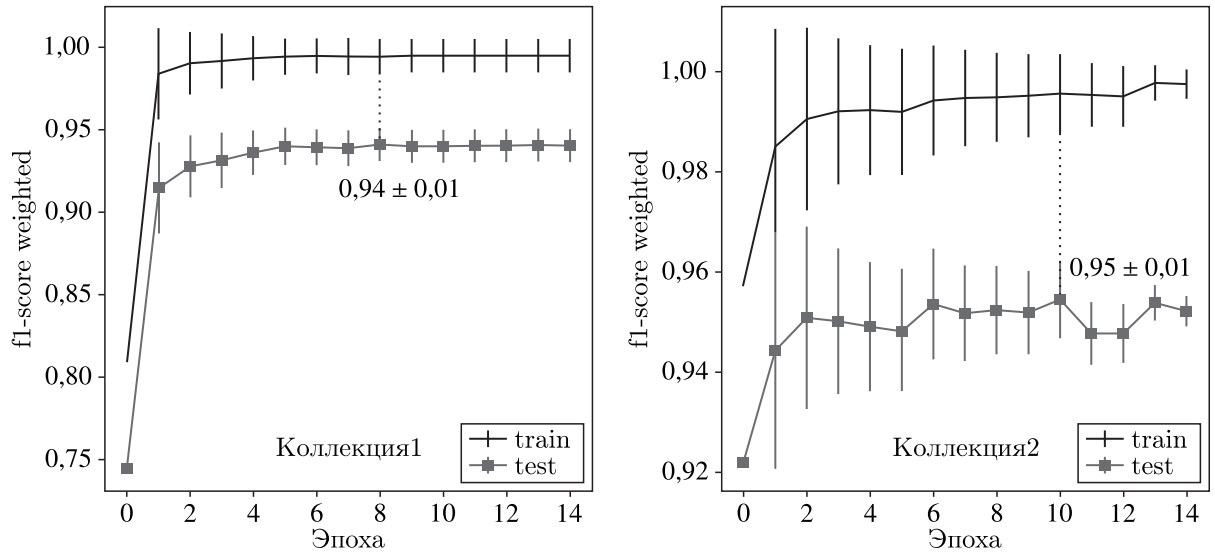

Рис. 6. Зависимость точности F1-score weighted классификатора MLPClassifier от номера эпохи для обучающей и отложенной выборок разных коллекций

Дополнительно метод контрастного семплирования для решения задачи предсказания библиографических ссылок был проверен на линейном классификаторе SGDClassifier [Scikit-learn], обучающемся с помощью стохастического градиентного спуска (кусочно-линейная функция потерь (hinge loss)). Балансировка обучающей выборки по классам производилась с помощью параметров классификатора, в котором предусмотрена автоматическая настройка весов в зависимости от частот классов во входных данных.

На рис. 7 показана зависимость точности F1-score weighted классификатора SGDClassifier от номера эпохи для обучающей и отложенной выборок на обеих коллекциях.

В рамках анализа качества работы метода контрастного семплирования был рассмотрен набор семплов, на которых классификатор делает неверный вывод, — примеры false positive и false negative из матрицы ошибок (confusion matrix). В таблице 3 приведены примеры семплов,

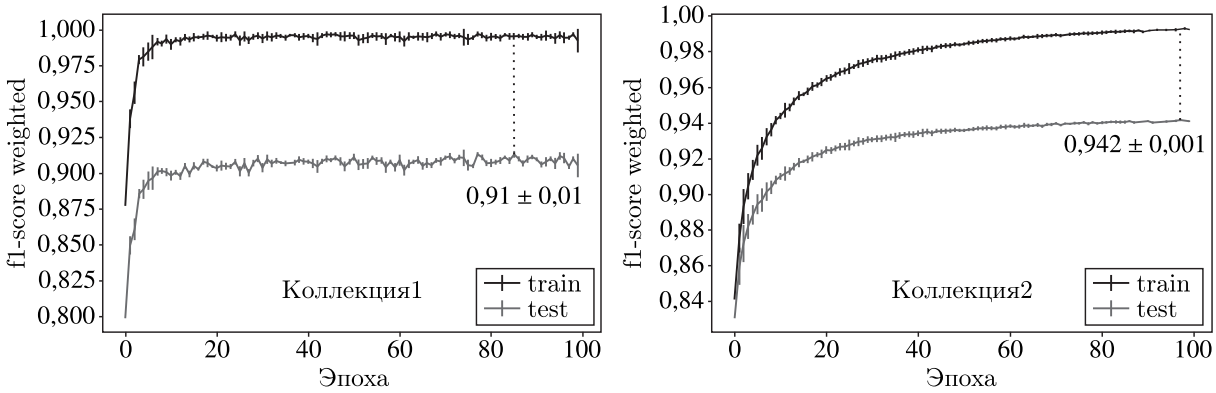

Рис. 7. Зависимость точности F1-score weighted классификатора SGDClassifier от номера эпохи для обучающей и отложенной выборок разных коллекций

для которых классификатор предсказал необходимость ссылки с высокой вероятностью, однако фактически ссылок во фрагментах нет; в табл. 4 приведены примеры семплов, для которых классификатор предсказал отсутствие ссылки, хотя она есть (примеры взяты из набора Коллекция1 из журнала «Правоприменение»).

Таблица 3. Примеры негативных семплов, для которых с высокой вероятностью предсказана ссылка

№	Содержимое семпла	Вероятность ссылки
1	По мнению суда, к государственным относятся функции, реализация которых обеспечивает общественное благосостояние в целом. Сюда относятся полицейские функции, пожарная безопасность и отправление правосудия в определяемых государством границах. Осуществление же муниципальных функций, по мнению суда, «приносит пользу непосредственно сообществу граждан, проживающих в пределах соответствующей территории».	0,98
2	Развитие права нашей страны связано с множеством факторов. Одним из них является изучение и адаптирование опыта других государств в регулировании различных правоотношений. В данном контексте особенно интересна практика так называемых судов справедливости.	0,98
3	Отметим, что и политическая воля может испытывать влияние внешних факторов (в демократических государствах в роли такого фактора часто выступает общественное мнение). Российская же правотворческая и правореализационная практика зачастую на словах свободна от таких факторов, что вызывает как обоснованные претензии со стороны общественности, так и критику ученых. Декларативно закрепленные возможности науки влиять на правотворческую практику далеко не всегда реально учитываются законодателем в правотворческой деятельности.	0,97

В таблице 3 приведены примеры, когда рекомендация классификатора кажется вполне оправданной, поскольку во фрагментах содержатся цитаты и отсылки к выражению мнения (претензии, критика), что требует подтверждения в виде ссылки на источник. В таблице 4 показаны фрагменты, когда необходимость ссылки именно в данном месте текста не является очевидной, что совпадает с выводом классификатора. Следовательно, часть ошибок из confusion matrix на самом деле не являются таковыми, эти случаи можно трактовать как указание на возможную неточность при составлении текста.

Таким образом, метод контекстного семплирования подтвердил свою состоятельность и может быть использован авторами (или редакторами) для подтверждения целесообразности

Таблица 4. Примеры позитивных семплов, для которых ссылка предсказана с низкой вероятностью

№	Содержимое семпла	Вероятность ссылки
1	К мерам натуральной помощи, как правило, относятся: бесплатный проезд учащихся общеобразовательных школ на внутригородском транспорте и в автобусах пригородных и внутрирайонных линий; первоочередной прием детей в дошкольные учреждения; бесплатные завтраки и обеды для учащихся общеобразовательных и профессиональных учебных заведений; бесплатное обеспечение школьной и спортивной формой на период обучения в общеобразовательной школе; бесплатное посещение музеев, парков, выставок (один день в месяц); первоочередное выделение садово-огородных участков; бесплатная выдача лекарств (по рецептам врачей) для детей в возрасте до шести лет [ссылка]. Предусматривается также содействие в предоставлении льготных кредитов, дотаций, беспроцентных ссуд на приобретение строительных материалов и строительство жилья; учет необходимости трудоустройства многодетных родителей при разработке региональных программ занятости, возможности их работы на условиях применения гибких форм труда; организация обучения и переобучения многодетных родителей с учетом потребностей экономики региона. Региональным законодательством устанавливаются меры социальной поддержки молодых семей.	$2 \cdot 10^{-11}$
2	Согласно статье 37 Федерального закона «О службе в органах внутренних дел Российской Федерации и внесении изменений в отдельные законодательные акты Российской Федерации»2 служба в органах внутренних дел приостанавливается в случае: — назначения (избрания) сотрудника органов внутренних дел на государственную должность Российской Федерации, членом Совета Федерации Федерального Собрания Российской Федерации или депутатом Государственной Думы Федерального Собрания Российской Федерации, депутатом законодательного (представительного) органа государственной власти субъекта Российской Федерации, 2 Собрание законодательства Российской Федерации. Аналогичные случаи предусмотрены статьей 43.1 Федерального закона «О прокуратуре Российской Федерации»3, статьей 45 Федерального закона «О воинской обязанности и военной службе»4. В статье 348.4 ТК РФ5 законодатель попытался определить правовую сущность приостановления трудового договора [ссылка].	10^{-7}
3	В связи с этим считаем, что отчасти корень проблемы нарушения принципов уголовного права, которые должны оказывать влияние как на правотворческую, так и на правоприменительную деятельность в ходе реализации различных мер уголовноправового воздействия, определяя специфику каждой из них [ссылка], кроется в искусственно созданной конкуренции норм и, как следствие, размытии пределов регулирования. Такое решение законодателя не является (к сожалению) единичным. Выделение стадий совершения преступления или форм соучастия в самостоятельные уголовно-наказуемые деяния встречается в главе о преступлениях против общественной безопасности.	0,004

ссылки и/или для обнаружения недостающих ссылок в процессе обеспечения целостности повествования. Также было обнаружено, что точность может быть повышена за счет еще большего расширения набора признаков, характерных для предметной области публикации, а также учета структурных элементов статьи — границ абзацев и разделов, анализа ссылок в конкретных разделах публикаций.

Результаты

Сформулированные исследовательские гипотезы подтвердились экспериментально. В базовом варианте алгоритма, позволяющего обнаруживать недостающие ссылки в тексте научной

статьи с помощью автоматической бинарной классификации, точность классификации составляет порядка 70 %, что не является достаточным для практического применения в условиях реальных информационных систем. Улучшение результатов было достигнуто благодаря следующим модификациям: контрастное семплирование при формировании обучающей выборки; изменение пространства признаков с помощью обработки именованных сущностей в тексте. На исследуемых данных (русскоязычные научные статьи по юридической и медицинской тематике) дополненный алгоритм показал точность 95 %, что является сравнимым с наилучшими результатами работ по решению задачи NER с помощью автоматической классификации: достигнутый в настоящее время максимум точности, описанный в научных работах, составляет 94,6 % [ExplainaBoard].

Представленный алгоритм имеет следующие ограничения.

- Проверка предложенного алгоритма на других (опубликованных) корпусах не проводилась. Однако, по мнению авторов, она не является целесообразной в силу построения набора признаков, при котором использовались экспертные знания о сущности «ссылка» в предметной области, то есть в текстах научных статей.

- При использовании полученного алгоритма рассматривались ссылки на научные работы. Для определения необходимости ссылок на НПА в юридических текстах может потребоваться модификация набора признаков с учетом особенностей принятых норм и правил цитирования законодательства.

- Учитывались не все сущности, информация о которых может быть дополнительно добавлена в набор признаков; например, не выделялись временные интервалы, отсылки к определенным датам (годам).

- Связь семплов со структурными элементами документа не анализировалась, хотя дополнительно возможно рассматривать семплы в привязке к разделу документа, исключать из рассмотрения заголовки и подписи к иллюстрациям, ввести запрет на формирование семпла из предложений разных абзацев.

Проведенные эксперименты показали, что на точность решения существенное влияние оказывают одновременно и выбор метода семплирования текста, и построение набора признаков с учетом экспертных знаний об именованной сущности в структуре исследуемого текста. При этом метод получения семплов текста может быть отдельно подвергнут более тонкой донастройке, например: усиление требования контрастности по отношению к негативным семплам (расширение исходного фрагмента без ссылок), увеличение шага при сдвиге скользящего окна (с одного до двух предложений).

Следует отметить, что, хотя наиболее высокую точность показал нейросетевой классификатор (95 %), по скорости работы он существенно уступает линейному классификатору (точность которого 91–94 %), однако при этом их точность можно считать сопоставимой. В случае если на практике достаточной является точность чуть более 90 %, то алгоритм классификации на основе линейной регрессии может быть реализован в прикладной системе. При этом обработка документов может происходить в онлайн-режиме, поскольку разметка производится автоматически и также не является ресурсоемкой.

Как было сказано в вводной части статьи, можно рассматривать задачу определения недостающих или лишних ссылок в тексте как вариант задачи определения тональности с двумя полярными классами: искомой тональностью является «неуверенность» автора, необходимость подкрепления/обоснования сформулированного утверждения. При таком подходе полученные в данной работе результаты также превосходят существующие: максимум точности на текстовых данных со сравнимыми характеристиками в настоящее время находится на уровне 70–80 % [Sentiment].

Таким образом, точность решения в исследуемой прикладной задаче зависит не только от метода классификации и его параметров, но также и от подготовки данных с учетом всех доступных знаний об обрабатываемых текстах. В этом случае затраты на формирование оптимальных семплов и признаков компенсируются упрощением вычислений: в частности, в данном исследовании достаточным методом векторизации оказался подсчет встречаемости слов в модели BoW, тогда как более затратный метод построения векторного пространства TF-IDF не дал существенного улучшения результата.

Заключение

В работе предложен новый метод, позволяющий анализировать вероятность библиографической ссылки во фрагменте научной статьи, — контрастное семплирование. Постановка задачи, сделанная авторами, развивает уже хорошо исследованные области анализа текстов для выявления именованных сущностей и тональности текста и одновременно является основой для выделения нового класса задач (определения оптимального размера контекста при анализе текста). Основным техническим новшеством предлагаемого метода является понятие контекста ссылки — границ фрагмента текста, максимально влияющего на вероятность нахождения в нем библиографической ссылки. В большинстве существующих методик анализа текста размеру фрагмента не придается существенного значения, но в данной статье показана критическая важность как размера фрагмента, так и окружения этого фрагмента. Фрагменты, которые становятся семплами разных классов в обучающей выборке, должны обладать свойством контрастности. Образно можно говорить, что контрастность фрагмента контекста ссылки показывает, что смыслы, относящиеся к библиографической ссылке, локализованы в данном фрагменте. Важно отметить высокую вычислительную эффективность предложенного метода по сравнению со сверточными искусственными нейронными сетями из-за значительных размеров фрагментов. Исследованный подход к анализу текстов расширяет методику Bidirectional Encoder Representations from Transformers (BERT), нацеленную на обучение модели влиянию глобального и локального контекста на языковую модель.

Список литературы (References)

Акоев М. А. и др. Руководство по наукометрии: индикаторы развития науки и технологии. — 2014.
Akoev M. A. et al. Rukovodstvo po naukometrii: indikatori razvitiia nauki i tehnologii [Handbook for Scientometrics: Indicators of science and technology development]. — 2014 (in Russian).

Носовец С. Г. Гипертекстовые ссылки в интернет-СМИ: опыт типологической характеристики // Вестник Челябинского государственного университета. — 2011. — № 17.
Nosovec S. G. Gipertekstovye ssylki v internet-SMI: opyt tipologicheskoj harakteristiki [Hypertext links in Internet media: experience of typological characteristics] // Bulletin of Chelyabinsk State University. — 2011. — No. 17 (in Russian).

Проект Natasha — набор Python-библиотек для обработки текстов на естественном русском языке. — [Электронный ресурс]. — URL: https://natasha.github.io/ (дата обращения: 29.07.2021).
Proekt Natasha — nabor Python-bibliotek dlya obrabotki tekstov na estestvennom russkom yazyke [Natasha project — a set of Python libraries for natural Russian word processing]. — [Electronic resource]. — Available at: https://natasha.github.io/ (accessed: 29.07.2021) (in Russian).

Стройков С. А. Основные понятия лингвистической концепции электронного лексикографического гипертекста // Известия Самарского научного центра Российской академии наук. — 2010. — Т. 12, № 5-3.
Strojkov S. A. Osnovnye ponyatiya lingvisticheskoj koncepcii elektronnogo leksikograficheskogo giperteksta [Basic Notions of the Linguistic Conception of Electronic Lexicographical Hypertext] // Izvestiya Samarskogo nauchnogo tsentra RAN. — 2010. — Vol. 12, no. 5 (3) (in Russian).

Aljuaid H. et al. Important citation identification using sentiment analysis of in-text citations // Telematics and Informatics. — 2021. — Vol. 56. — P. 101492.

Apache Tika. — [Electronic resource]. — URL: https://tika.apache.org/ (accessed: 29.07.2021).

Arsyad S. et al. The rhetorical problems experienced by Indonesian lecturers in social sciences and humanities in writing research articles for international journals // The Asian Journal of Applied Linguistics. — 2020. — Vol. 7, no. 1. — P. 116–129.

Bojanowski P. et al. Enriching word vectors with subword information // Transactions of the Association for Computational Linguistics. — 2017. — Vol. 5. — P. 135–146.

Bornmann L., Wagner C., Leydesdorff L. The geography of references in elite articles: Which countries contribute to the archives of knowledge? // PloS one. — 2018. — Vol. 13, no. 3. — P. e0194805.

Bornmann L., Wray K. B., Haunschild R. Citation concept analysis (CCA): a new form of citation analysis revealing the usefulness of concepts for other researchers illustrated by exemplary case studies including classic books by Thomas S. Kuhn and Karl R. Popper // Scientometrics. — 2020. — Vol. 122, no. 2. — P. 1051–1074.

Brown T. B. et al. Language models are few-shot learners // arXiv preprint arXiv:2005.14165. — 2020.

Chandrasekaran M. K. et al. Overview and results: Cl-scisumm shared task 2019 // arXiv preprint arXiv:1907.09854. — 2019.

Chandrasekharan S. et al. Finding scientific communities in citation graphs: Articles and authors // Quantitative Science Studies. — 2021. — Vol. 2, no. 1. — P. 184–203.

Dernoncourt F., Lee J. Y., Szolovits P. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks // arXiv preprint arXiv:1705.05487. — 2017.

Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding // arXiv preprint arXiv:1810.04805. — 2018.

Emerson L., Rees M. T., MacKay B. Scaffolding academic integrity: Creating a learning context for teaching referencing skills // Journal of university teaching & learning practice. — 2005. — Vol. 2, no. 3. — P. 3.

ExplainaBoard — Named Entity Recognition. — [Electronic resource]. — URL: http://explainaboard.nlpedia.ai/leaderboard/task-ner/ (accessed: 29.07.2021).

Fu J., Huang X., Liu P. SpanNer: Named Entity Re-/Recognition as Span Prediction // arXiv preprint arXiv:2106.00641. — 2021.

Gallant S. I. A practical approach for representing context and for performing word sense disambiguation using neural networks // Neural Computation. — 1991. — Vol. 3, no. 3. — P. 293–309.

Gray K. et al. Web 2.0 authorship: Issues of referencing and citation for academic integrity // The Internet and Higher Education. — 2008. — Vol. 11, no. 2. — P. 112–118.

Herrmannova D., Knoth P. Simple yet effective methods for large-scale scholarly publication ranking // arXiv preprint arXiv:1611.05222. — 2016.

Huang E. H. et al. Improving word representations via global context and multiple word prototypes // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — 2012. — P. 873–882.

Huang R., Krylova K. Team MLU@ CL-SciSumm20: Methods for Computational Linguistics Scientific Citation Linkage // Proceedings of the First Workshop on Scholarly Document Processing. — 2020. — P. 282–287.

Joulin A. et al. Bag of tricks for efficient text classification // arXiv preprint arXiv:1607.01759. — 2016.

Kingma D. P., Ba J. Adam: A method for stochastic optimization // arXiv preprint arXiv:1412.6980. — 2014.

Klavans R., Boyack K. W. Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? // Journal of the Association for Information Science and Technology. — 2017. — Vol. 68, no. 4. — P. 984–998.

Lai K. K. et al. Identifying the impact of patent family on the patent trajectory: A case of thin film solar cells technological trajectories // Journal of Informetrics. — 2021. — Vol. 15, no. 2. — P. 101143.

Li B. Named entity recognition in the style of object detection // arXiv preprint arXiv:2101.11122. — 2021.

MacRoberts M. H., MacRoberts B. R. The mismeasure of science: Citation analysis // Journal of the Association for Information Science and Technology. — 2018. — Vol. 69, no. 3. — P. 474–482.

Manggala P. et al. On augmenting the references section with a citation network visualization // Beyond static papers: Rethinking how we share scientific understanding in ML-ICLR 2021 workshop. — 2021.

Medić Z., Snajder J. Improved Local Citation Recommendation Based on Context Enhanced with Global Information // Proceedings of the First Workshop on Scholarly Document Processing. — 2020. — P. 97–103.

Merton R. K. The sociology of science: Theoretical and empirical investigations. — University of Chicago press, 1973.

Mikolov T. et al. Distributed representations of words and phrases and their compositionality // Advances in neural information processing systems. — 2013. — P. 3111–3119.

Miura T., Asatani K., Sakata I. Classifying Sleeping Beauties and Princes Using Citation Rarity // International Conference on Complex Networks and Their Applications. — Springer, Cham, 2020. — P. 308–318.

Mozharova V., Loukachevitch N. Two-stage approach in Russian named entity recognition // 2016 International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT). — IEEE, 2016. — P. 1–6.

Pears R., Shields G. J. Cite them right: the essential referencing guide. — Macmillan International Higher Education, 2019.

Pennington J., Socher R., Manning C. D. Glove: Global vectors for word representation // Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). — 2014. — P. 1532–1543.

Prester J. et al. Classifying the ideational impact of information systems review articles: A content-enriched deep learning approach // Decision Support Systems. — 2021. — Vol. 140. — P. 113432.

Russian GPT-3 models. — [Electronic resource]. — URL: https://github.com/sberbank-ai/ru-gpts (accessed: 29.07.2021).

Scikit-learn. Machine Learning in Python. — [Electronic resource]. — URL: https://scikit-learn.org/ (accessed: 29.07.2021).

Sentiment Analysis in Russian. — [Electronic resource]. — URL: https://github.com/sismetanin/sentiment-analysis-in-russian (accessed: 29.07.2021).

Tahamtan I., Bornmann L. What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018 // Scientometrics. — 2019. — Vol. 121, no. 3. — P. 1635–1684.

Trujillo C. M., Long T. M. Document co-citation analysis to enhance transdisciplinary research // Science advances. — 2018. — Vol. 4, no. 1. — P. e1701130.

Valencia-Hernández D. S. et al. SAP Algorithm for Citation Analysis: An improvement to Tree of Science // Ingeniería e Investigación. — 2020. — Vol. 40, no. 1. — P. 45–49.

Wang X. et al. Improving Named Entity Recognition by External Context Retrieving and Cooperative Learning // arXiv preprint arXiv:2105.03654. — 2021.

Yang S., Wang F. Visualizing information science: Author direct citation analysis in China and around the world // Journal of Informetrics. — 2015. — Vol. 9, no. 1. — P. 208–225.

Ziyadi M. et al. Example-based named entity recognition // arXiv preprint arXiv:2008.10570. — 2020.