DOI: 10.15514/ISPRAS-2022-34(2)-7

Стратегии семплирования текста для прогнозирования недостающих библиографических ссылок

Ф.В. Краснов, ORCID: 0000-0002-9881-7371 <fkrasnov2@yandex.ru>
И.С. Смазневич, ORCID: 0000-0002-5996-4635 <ismaznevich@naumen.ru>
Е.Н. Баскакова, ORCID: 0000-0002-7071-8961 <enbaskakova@naumen.ru>
NAUMEN,
620028, Россия, Екатеринбург, ул. Татищева, 49А, БЦ «Татищевский», 4 этаж

Аннотация. В статье исследуются различные стратегии семплирования текстовых данных при выполнении автоматической классификации предложений с целью обнаружения недостающих библиографических ссылок. Построение семплов осуществляется на основе предложений в качестве семантических единиц текста, к которым добавляется их непосредственный контекст, состоящий из нескольких соседних предложений. Исследуется ряд стратегий семплирования, которые различаются размером и положением контекста. Эксперимент проведен на данных из сборника научных работ по естественнонаучной и инженерной тематике. Показано, что включение контекста предложений в семплы улучшает результат классификации предложений. Предложен метод автоматического определения оптимальной стратегии семплирования для данной текстовой коллекции: оптимальная стратегия определяется результатом голосования одинаковых классификаторов, получающих на вход одни и те же данные, семплированные различными способами. Семплирование с учетом контекста предложения в сочетании с процедурой жесткого голосования (hard voting) показало точность классификации 98% (оценка F1). Предложенный подход к обнаружению недостающих библиографических ссылок может использоваться в рекомендательных модулях прикладных интеллектуальных информационных систем.

Ключевые слова: семплирование текста; стратегия семплирования; анализ цитирования; прогнозирование библиографических ссылок; классификация предложений

Для цитирования: Краснов Ф.В., Смазневич И.С., Баскакова Е.Н. Стратегии семплирования текста для прогнозирования недостающих библиографических ссылок. Труды ИСП РАН, том 34, вып. 2, 2022 г., стр. 77-88. DOI: 10.15514/ISPRAS–2022–34(2)–7

Text sampling strategies for predicting missing bibliographic links

F.V. Krasnov, ORCID: 0000-0002-9881-7371 <fkrasnov2@yandex.ru>
I.S. Smaznevich, ORCID: 0000-0002-5996-4635 <ismaznevich@naumen.ru>
E.N. Baskakova, ORCID: 0000-0002-7071-8961 <enbaskakova@naumen.ru>
NAUMEN,
49A, Tatishcheva st., Yekaterinburg, 620028, Russia

Abstract. The paper proposes various strategies for sampling text data when performing automatic sentence classification for the purpose of detecting missing bibliographic links. We construct samples based on sentences as semantic units of the text and add their immediate context which consists of several neighbouring sentences. We examine a number of sampling strategies that differ in context size and position. The experiment is carried out on the collection of STEM scientific papers. Including the context of sentences into samples improves the result of their classification. We automatically determine the optimal sampling strategy for a given text collection by implementing an ensemble voting when classifying the same data sampled in different ways. Sampling strategy taking into account the sentence context with hard voting procedure leads to the classification accuracy of 98% (F1-score). This method of detecting missing bibliographic links can be used in recommendation engines of applied intelligent information systems. Keywords: text sampling, sampling strategy, citation analysis, bibliographic link prediction, sentence classification.

Keywords: text sampling; sampling strategy; citation analysis; prediction of bibliographic references; proposition classification

For citation: Krasnov F.V., Smaznevich I.S., Baskakova E.N. Text sampling strategies for predicting missing bibliographic links. Trudy ISP RAN/Proc. ISP RAS, vol. 34, issue 2, 2022, pp. 77-88 (in Russian). DOI: 10.15514/ISPRAS–2022–34(2)–7

1. Введение

Научное исследование невозможно без соотнесения полученных результатов с работами других ученых: их следует упомянуть, вставив в статью библиографические ссылки. Специалисты в области наукометрии по-разному обосновывают необходимость установления таких связей между исследованиями и формулируют различные теории цитирования.

Нормативная теория цитирования, которая опирается на принципы научной этики, сформулированные Мертоном (Robert K. Merton) [1], предполагает, что ссылки в научных статьях призваны указывать на работы, которые являются основой для исследования, связаны тематически, описывают используемые методы или необходимы для обсуждения результатов. Согласно рефлексивной теории, связи между научными работами указывают на состояние науки и помогают создать ее формализованное представление, например, карты науки [2-3].

Таким образом, бенефициаром корректности научного цитирования является все научное сообщество, как исследователи, создающие статьи о своих результатах, так и администраторы, отслеживающие достижения в различных областях науки. Упоминание актуальных и значимых результатов других ученых является одним из основных требований при построении научных текстов, в частности, с точки зрения редакторов научных журналов. Эти требования отмечены в руководствах по академическому письму [4-6] и подтверждаются на практике, что описано, например, в результатах исследований публикационной активности в высокорейтинговых международных журналах [7].

Авторы научных работ самостоятельно выбирают источники для цитирования и позиции для ссылок в тексте, и в настоящее время этот процесс никак не автоматизирован. В данной работе исследуется возможность создания рекомендательного алгоритма, позволяющего находить недостающие библиографические ссылки в научной статье, то есть выявлять те фрагменты текста, где необходимо упомянуть другую исследовательскую работу. Для этой цели оценивается вероятность наличия ссылок во фрагментах текста, используя обучение с частичным привлечением учителя. Формальная постановка рассматриваемой задачи заключается в следующем: требуется автоматически находить в тексте научной статьи те фрагменты (предложения), где ссылка отсутствует, но необходима, используя в качестве обучающих данных набор размеченных фрагментов со ссылками и без ссылок.

Задача классификации фрагментов текста в зависимости от наличия в них ссылок методологически схожа с задачей анализа тональности, в рамках которой тексты автоматически классифицируются как позитивные и негативные (в основном) в соответствии с их эмоциональными характеристиками. В дополнение к классификации фрагментов на позитивные и негативные принцип анализа тональности используется для выделения других классов, включая определение значимости цитирования [8-11]. Задача выявления недостающих или ненужных ссылок в тексте также может рассматриваться аналогично анализу тональности, тогда искомым настроением здесь является потребность автора подтвердить сформулированное утверждение.

Другим близким направлением исследований является распознавание именованных сущностей (Named Entity Recognition, NER) на основе предсказания классификатора. Аналогичная задача рассматривается в работе [12], где сущности выделяются через определение спанов (span prediction). Задача NER может быть решена в два этапа: идентификация фрагментов с высокой вероятностью содержания сущностей и определение точного положения этих сущностей [13-14]. Некоторые методы NER также учитывают контекст сущностей, как локальный, так и глобальный, а также внешний [15].

Задача классификации предложений с учетом их ближайшего контекста обсуждалась в ряде исследований. В работе [16] использовались контекстно-зависимые эмбеддинги, созданные языковыми моделями, высокое качество которых достигается ценой скорости. Автором [17] изучалась тематическая классификация и было показано, что модели, получающие контекст среди входных данных, работают лучше моделей без контекста. В этих работах размер контекста определяется единожды исходя из некоторых предположений и для конкретного текстового корпуса он может оказаться неоптимальным.

Метод, представленный в данной работе, также можно рассматривать как своего рода ресемплирование (англ. resampling). До сих пор оно выполнялось в основном для балансировки распределения классов в обучающей выборке с целью повышения точности предсказания, на которую несбалансированность данных влияет негативно. Методы балансировки подразделяются на три типа, а именно: сокращение количества объектов мажоритарного класса (undersampling), увеличение количества объектов миноритарного класса (oversampling) и гибридные методы. Первый подход подразумевает исключение некоторых данных мажоритарного класса (см., например, [18-19]); второй предполагает воспроизведение существующих экземпляров миноритарного класса либо создание новых [20-22], а гибридные методы направлены на объединение преимуществ обоих подходов [23].

Локальные и глобальные контексты учитываются современными архитектурами нейронных сетей для анализа текста. Поскольку текст представляет собой однонаправленный список термов, контекст обычно понимается как некоторая совокупность соседних слов до или после рассматриваемого терма [24-25]. В сверточных нейронных сетях увеличение размера контекста приводит к значительному увеличению размерности тензоров и, как следствие, количества параметров модели, что в свою очередь требует увеличения размера коллекций. В моделях глубокого обучения типа transformer контекст учитывается с помощью механизма внимания, а локальный контекст сочетается с более широким контекстом (BERT [26], GPT-3 [27]).

Важно, что все вышеперечисленные алгоритмы не учитывают естественные структурные единицы текстов (предложения и абзацы), поскольку эти алгоритмы настраиваются на определенный размер контекста, составляющий фиксированное количество слов, в то время как размер предложений и абзацев варьируется.

2. Методы

Задача определения недостающих ссылок формализуется как поиск фрагментов текста, где ссылка отсутствует, но необходима, или, наоборот, присутствует, но не нужна.

Решается задача автоматической классификации с двумя классами (позитивным и негативным). Для каждого фрагмента научной статьи предложенный алгоритм определяет вероятность наличия в нем библиографической ссылки. Набор текстовых документов задается так, что каждый документ состоит из фрагментов. Фрагмент представляет собой последовательность слов (термов) разной длины. Фрагменты могут накладываться друг на друга и различаться по размеру. Каждый фрагмент представляет собой семпл и помечается меткой одного из двух возможных классов: позитивного или негативного. Метка класса соответствует тому, содержит ли данный фрагмент библиографическую ссылку или нет. Задачей данного исследования является поиск такой стратегии построения фрагментов

(семплирования), которая дает наибольшую точность в определении меток класса для заданного классификатора.

Гипотеза исследования заключается в следующем: стратегии семплирования текста, учитывающие контекст, повышают точность классификации предложений, используемой для прогнозирования недостающих библиографических ссылок в научных статьях.

Позитивный семпл состоит из библиографической ссылки, окруженной ее контекстом из исходного текста, а негативный семпл представляет собой фрагмент без библиографической ссылки в нем. Чтобы избежать дублирования образцов, предложение с двумя или более ссылками рассматривается только один раз. Контекст ссылки ограничивается содержащим ее предложением либо расширяется и включает в себя также соседние предложения.

Наилучшим вариантом является ситуация, когда границы контекста ссылки совпадают с границами законченной мысли автора, к которой относится эта ссылка. В этом случае смысловой единицей текста могут быть как одно, так и несколько предложений, что затрудняет определение размера контекста. Тем не менее, чтобы приблизиться к наилучшей ситуации, в предложенном алгоритме в качестве контекста рассматривается фрагмент, размер которого определяется количеством предложений, а не слов (в отличие от алгоритмов нейронной сети). Таким образом, контекст формируется на основе естественных структурных единиц текста.

Пространство признаков создается автоматически на основе статистики словаря в рамках модели «Мешок слов» (Bag of Words). Словарь модели включает в себя слова и все оригинальные знаки препинания и служебные символы. В качестве дополнительных признаков рассматриваются именованные сущности.

3. Алгоритм

Алгоритм состоит из следующих этапов.

- Предварительная обработка текста.
- Очистка текста: удаление служебных символов (табуляция, перевод строки и т. д.), слов (названия журналов, ISBN и т.д.) и разделов (информация о финансировании, список литературы);
- Токенизация:
  - разбиение текста на предложения;
  - нормализация термов.
- Разметка данных.
- Для каждого документа (статьи) добавляются метки начала и конца;
- Для каждого предложения:
  - если предложение содержит библиографическую ссылку (обозначение цитирования), оно помечается как относящееся к классу "Со ссылками";
  - если в предложении нет обозначения цитирования, оно получает метку класса "Без ссылок".
- После разметки предложений обозначения цитирования удаляются.
- Обработка именованных сущностей.
- Обнаружение именованных сущностей в тексте;
- Замена именованных сущностей специальными метками.
- Конструирование семплов. Семплы строятся по-разному в зависимости от класса (позитивного или негативного):
  - Позитивный семпл строится из одного предложения “Со ссылками”, к которому добавляется n предыдущих предложений и m последующих предложений; все

предложения берутся в исходном порядке.

- Негативный семпл составляется из k предложений "Без ссылок", идущих подряд в исходном тексте (смежных предложений), где k = n + 1 + m.

Структура семплов показана на рис. 1.

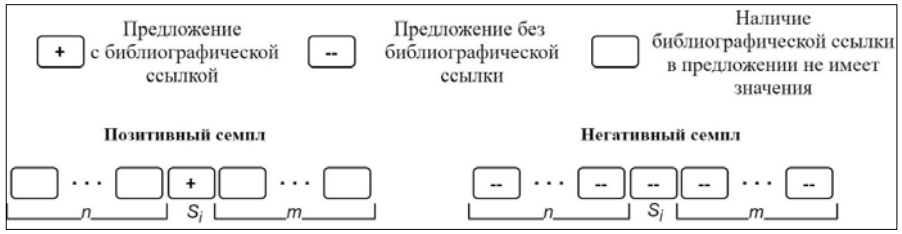

Рис. 1. Построение семплов
Fig. 1. Construction of samples

- Классификация семплов.
- Балансировка распределения классов в обучающей выборке выполняется методом случайного сокращения мажоритарного класса (random undersampling).
- Для каждого семпла строится векторная модель с использованием счетчика слов (как наиболее быстрого и эффективного с вычислительной точки зрения представления текста).
- Векторизованный набор семплов обрабатывается с помощью классификатора.
- Определение оптимальной стратегии конструирования семплов.

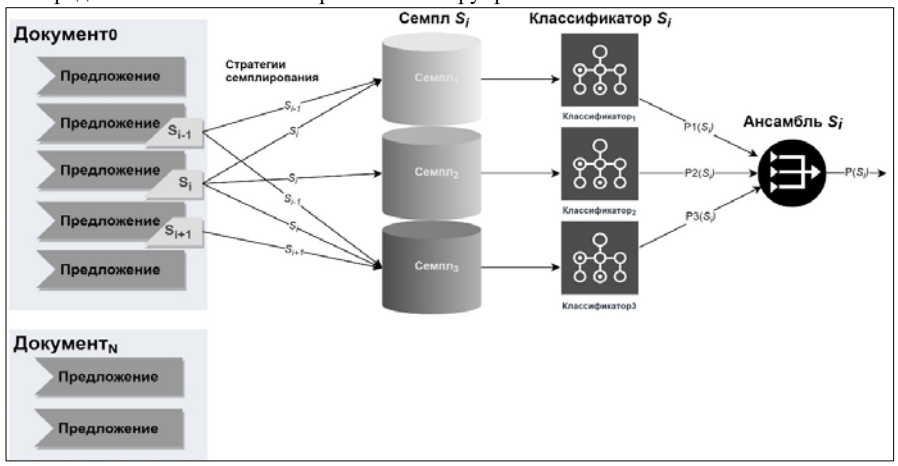

Рис. 2. Общая схема алгоритма классификации предложений на классы "Со ссылками" и "Без ссылок" с использованием различных стратегий семплирования (для n = m = 1)
Fig. 2. The algorithm flowchart for classifying sentences into the classes "With links" and "Without links" using various sampling strategies (for n = m = 1)

Далее осуществляется построение ансамбля моделей классификации с целью автоматического определения оптимальной стратегии семплирования, а именно: одни и те же данные, собранные в семплы разными способами, обрабатываются однотипными классификаторами, после чего реализуется процедура голосования.

Общая схема алгоритма показана на рис. 2. Каждый семпл для предложения S_i соответствует одной стратегии семплирования, и для каждой стратегии семплирования выполняется своя процедура классификации. При этом все классификаторы используют один и тот же метод, но в качестве входных данных используют разные типы семплов.

4. Эксперимент

Экспериментальная проверка гипотезы была проведена на коллекции научных статей по естественнонаучной и инженерной тематике (STEM) из базы arXiv.org [28]. Документы этого набора данных содержат только тексты, рисунки и таблицы удалены. Математические формулы и обозначения цитирования заменены специальными токенами – @xmath<число> и @xcite (метки цитирования). Документы содержат только разделы до «Заключения» включительно, все последующие разделы удалены.

Размер набора данных следующий: количество документов – 215 тыс., средняя длина документа – 4938 слов, средняя длина аннотации – 220 слов.

Файлы представлены в формате jsonlines, где каждая строка представляет собой объект json, соответствующий одной научной статье. Каждая строка содержит аннотацию, перечень наименований разделов и основную часть статьи, где весь текст разделен на предложения.

В эксперименте рассматриваются предложения длиной более 30 слов. С учетом этого ограничения набор данных состоит в общей сложности из 458774 предложений.

Предложения, содержащие специальные метки цитирования @xcite, относятся к позитивному классу ("Со ссылками"), после чего метки цитирования удаляются. Предложения без меток цитирования классифицируются как негативные ("Без ссылок"). Соотношение классов следующее: 24% – предложения из позитивного класса, 76% – предложения из негативного класса. Такая разметка предложений считается нулевой стратегией семплирования (№0), и результат классификации данных, семплированных таким образом, рассматривается как базовый уровень: точность классификации со стратегией семплирования № 0, измеренная с помощью метрики F1, составляет 0,7866.

После установления базового уровня точности были протестированы различные стратегии семплирования данных с целью повышения точности классификации. Основная идея семплирования состоит в том, чтобы учитывать некоторый контекст предложений со ссылкой. Различные стратегии семплирования предполагают различные направления, позиции и размер контекста, определяемый числом соседних предложений. Каждое предложение [*i*] в различных стратегиях семплирования включается в разные типы (варианты) семплов.

В эксперименте тестировались 10 стратегий со следующими параметрами: *n*: [0, 1, 2 3 4 5], *m*: [0, 1, 2, 3, 4], *k*: [1, 3]. Все типы семплов, соответствующие выбранным стратегиям семплирования, представлены в табл. 1.

Табл. 1. Стратегии семплирования, протестированные в эксперименте
Table 1. Sampling strategies tested in the experiment

№	Стратегия семплирования (алгоритм построение семплов)	Количество предложений в семпле	
		Позитивный семпл	Негативный семпл
0	Предложение[*i*]	2640	12352
1	Предложение[*i*: *i* +2]	2640	10639
2	Предложение[*i* -1: *i*+1]	2640	10639
3	Предложение[*i* -1: *i*+2]	2640	9376
4	Предложение[*i* -2: *i*+2]	2640	8384
5	Предложение[*i* -3: *i*+2]	2640	7574
6	Предложение[*i* -3: *i*+3]	2640	6880
7	Предложение[*i* -4: *i*+3]	2640	6287
8	Предложение[*i* -4: *i*+4]	2640	5776

9	Предложение[*i* -5: *i*+4]	2640	5326

Распределение длины (количества слов) в позитивных и негативных семплах разных типов показано на рис. 3.

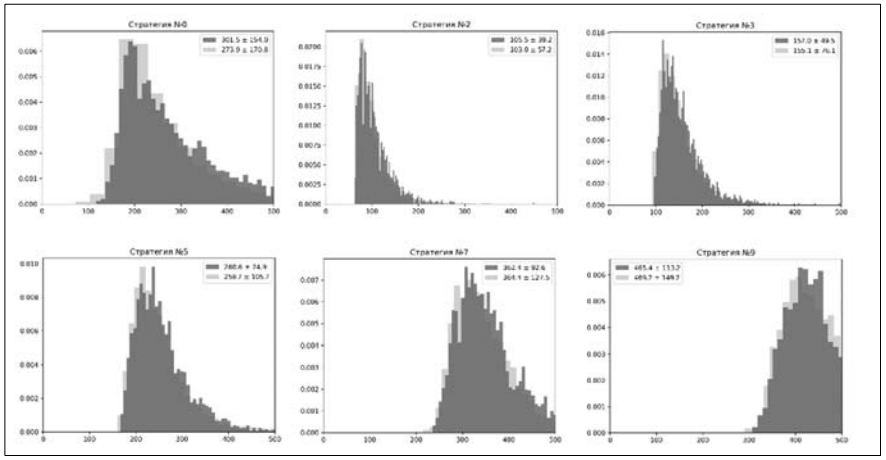

Рис. 3. Распределение количества слов в позитивных и негативных семплах разных типов (темно-серым цветом показаны позитивные семплы, светло-серым – негативные)
Fig. 3. The distribution of the length as a number of words in positive and negative samples of different types ('dark grey' color refers to the positive class, 'light grey' to the negative one)

После балансировки классов методом random undersampling данные были поделены на обучающую и тестовую выборки с параметром test_size=0.33.

Векторное представление строится с использованием метода CountVectorizer библиотеки Scikit-learn. Словарь включает в себя униграммы и биграммы и ограничен по частоте с параметрами min_df=3 (минимальная частота слова для исключения из словаря), max_df=0,7 (порог для исключения из словаря частовстречающихся слов по относительной частоте)

Для классификации используется многослойный персептрон (метод MLPClassifier библиотеки Scikit-learn). Эффективность классификации в зависимости от используемой стратегии семплирования представлена в табл. 2.

Для каждого предложения сравниваются результаты классификации, полученные при различных стратегиях семплирования, и дополнительно улучшаются с помощью процедуры голосования. Тестировались поочередно методы мягкого голосования (soft voting) и жесткого голосования (hard voting). При мягком голосовании для средней предсказанной вероятности было установлено пороговое значение 0,5. При жестком голосовании сумма всех предсказанных значений вероятности сравнивалась с пороговым значением 3. Было протестировано разное количество классификаторов, участвующих в голосовании. Комбинации классификаторов, соответствующих различным стратегия семплирования, формировались начиная с группы из трех классификаторов: № 7, № 8 и № 9, а затем к ним добавлялись еще классификаторы – один за другим в обратном порядке. Результаты голосования в зависимости от количества рассмотренных классификаторов показаны на рис. 4.

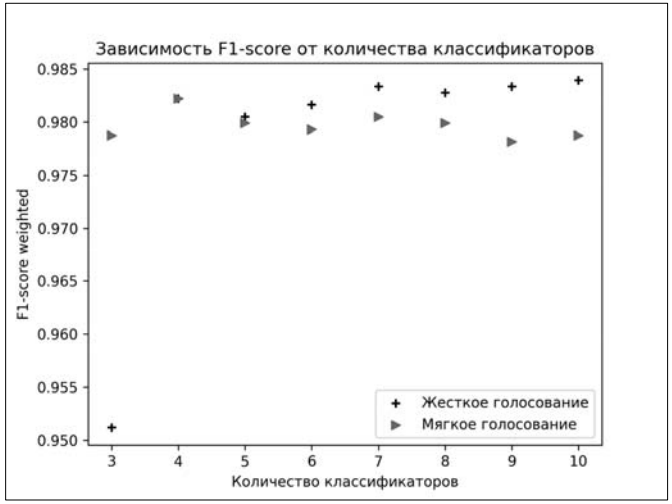

Рис. 4. Результат классификации с жестким и мягким голосованием в зависимости от количества участвующих классификаторов
Fig. 4. The result of classification with hard and soft voting depending on the number of estimators included

3. Результаты и обсуждение

Сформулированная исследовательская гипотеза была подтверждена экспериментально. Исследование показало, что выбор стратегии семплирования влияет на результат классификации текста.

Базовый уровень точности был установлен при использовании стратегии семплирования №0. В этом случае эффективность классификации, измеренная с помощью F1-меры, составляет всего 79%, что не является достаточным для практического использования в прикладных информационных системах.

Табл. 2. Результат классификации предложений классификатором MLP в зависимости от стратегии семплирования (средневзвешенное значение)
Table 2. The result of the classification of sentences by the MLP classifier depending on the sampling strategy (weighted average)

Стратегия семплирования	F1	Precision	Recall
0	0.7866	0.7866	0.7866
1	0.8882	0.8881	0.8881
2	0.8884	0.8881	0.8881
3	0.9214	0.9214	0.9214
4	0.9444	0.9443	0.9443
5	0.9410	0.9409	0.9409
6	0.9601	0.9598	0.9598
7	**0.9640**	**0.9639**	**0.9639**
8	0.9593	0.9593	0.9593
9	0.9581	0.9581	0.9581

Различные стратегии семплирования повышают эффективность классификации. В табл. 2 показано, что наилучший результат (оценка F1 96%) достигается при использовании стратегии семплирования № 7. Последующее увеличение количества предложений в семпле существенно не повышает точность, поскольку она стремится к асимптоте.

Дальнейшее улучшение достигается за счет стратегии семплирования данных, которая предполагает автоматическое определение оптимального типа семпла. Это обеспечивается

применением процедуры голосования к решениям, принятым различными классификаторами.

На рис. 4 показано, что процедура голосования дополнительно улучшает результаты классификации и увеличивает оценку F1 на 1,5%. При всех протестированных комбинациях классификаторов был получен стабильно высокий результат, но наилучшие значения достигались при жестком голосовании 7, 8 или 10 классификаторов, обрабатывающих длинные семплы.

Предложенный алгоритм показывает точность 98% (F1-мера), что сопоставимо с современными результатами для задачи NER с использованием автоматической классификации и других задач классификации текста [29]. Важно, что предложенный алгоритм обеспечивает высокую точность, но не требует для реализации огромных вычислительных ресурсов.

5. Заключение

В статье предлагается новый метод определения вероятности наличия библиографической ссылки во фрагментах научной статьи. Подход предполагает классификацию предложений с применением процедуры голосования, в котором различные стратегии семплирования данных используются классификаторами, реализующими один и тот же метод классификации. Постановка проблемы, сделанная авторами, близка к хорошо изученным задачам NER и анализа тональности, но является новой с точки зрения реального применения.

Основным новшеством предлагаемого метода является нахождение контекста ссылки, который максимально влияет на вероятность обнаружения недостающей библиографической ссылки в предложении. В предлагаемом алгоритме наилучший размер и положение контекста определяются автоматически. Размер определяется границами семантических единиц текста и измеряется количеством предложений, а не слов, таким образом используется тот факт, что предложение является более семантически емкой (значимой) единицей, чем слово. В большинстве существующих методов классификации текстов не предполагается, что контекст фрагмента имеет существенное значение, но данное исследование показывает критическую важность его рассмотрения. Значительное влияние контекста на эффективность классификации демонстрирует, что семантика, связанная с библиографической ссылкой, может быть локализована во фрагментах разной длины.

Точность предложенного алгоритма достигает 98% (оценка F1). Важно отметить высокую вычислительную эффективность описанного метода по сравнению со сверточными искусственными нейронными сетями. Это преимущество достигается за счет большего размера семплов. Исследуемый подход к анализу текста расширяет принцип механизма внимания, направленного на обучение языковой модели пониманию влияния глобального и локального контекстов. Автоматическое определение границ контекста коррелирует с идеей автоматического выбора значимых признаков в искусственных нейронных сетях.

Предлагаемый способ может быть использован в рекомендательных механизмах в прикладных интеллектуальных информационных системах, включая помощь в создании документов и составлении текстов с возможными ссылками на другие документы или помощь в проверке правильности документа. Такие функции полезны во многих областях, например, в науке, юриспруденции или журналистике, где документы содержат утверждения, которые должны быть подтверждены ссылками на правовые акты или другие источники.

Список литературы / References

[1] Merton R.K. The sociology of science: Theoretical and empirical investigations. University of Chicago press, 1973, 605 p.

[2] Москалева О.В., Акоев М.А. Наукометрия: немного истории и современные российские реалии. Управление наукой: теория и практика, том 1, no. 1, 2019 г., стр. 135-148 / Moskaleva O.V., Akoev M.A. Scientomentrics: a little bit of history and modern Russian realities. Science Management: Theory and Practice, vol. 1, no. 1, pp. 135-148 (in Russian).

[3] Зеленков Ю.А., Анисичкина Е.А. Динамика исследований в области интеллектуального анализа данных: тематический анализ публикаций за 20 лет. Бизнес-информатика, том 15, no. 1, 2021 г., стр. 30-46 / Zelenkov Yu.A., Anisichkina E.A. Trends in data mining research: A two-decade review using topic analysis. Business Informatics, vol. 15, no 1, 2021, pp. 30-46 (in Russian).

[4] Emerson L., Rees M. T., MacKay B. Scaffolding academic integrity: Creating a learning context for teaching referencing skills. Journal of university teaching & learning practice, vol. 2, issue 3, 2005, pp. 17-30.

[5] Gray K., Thompson C. et al. Web 2.0 authorship: Issues of referencing and citation for academic integrity. Internet and Higher Education. vol. 11, issue 2, 2008, pp. 112-118.

[6] Pears R., Shields G. Cite them right: the essential reference guide. Palgrave Macmillan, 8th edition, 2010, 112 p.

[7] Arsyad S., Ramadhan S., Maisarah I. 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, vol. 7, issue 1, 2020, pp. 116-129.

[8] Aljuaid H., Iftikhar R. et al. Important citation identification using sentiment analysis of in-text citations. Telematics and Informatics, vol. 56, 2021, article no. 101492.

[9] Prester J., Wagner G. et al. Classifying the ideational impact of information systems review articles: A content-enriched deep learning approach. Decision Support Systems, vol. 140, 2021, article no. 113432.

[10] Varanasi K.K., Ghosal T. et al. Iitp-cuni@ 3c: Supervised approaches for citation classification (task a) and citation significance detection (task b). In Proc. of the Second Workshop on Scholarly Document Processing, 2021, pp. 140-145.

[11] Färber M., Sampath A. Determining how citations are used in citation contexts. Lecture Notes in Computer Science, vol. 11799, 2019, pp. 380-383.

[12] Fu J., Huang X., Liu P. Spanner: Named entity re-/recognition as span prediction. arXiv.2106.00641, 2021, 13 p.

[13] Ziyadi M., Sun Y. et al. Example-based named entity recognition. arXiv.2008.10570, 2020, 15 p.

[14] Li B. Named entity recognition in the style of object detection. arXiv.2101.11122, 2021, 9 p.

[15] Wang X., Jiang Y. et al. Improving named entity recognition by external context retrieving and cooperative learning, arXiv.2105.03654, 2021, 13 p.

[16] Fiok K., Karwowski W. et al. Comparing the quality and speed of sentence classification with modern language models. Applied Sciences, vol. 10, issue 10, 2020, article no. 3386.

[17] Глазкова А.В. Тематическая классификация текстовых фрагментов с учетом их ближайшего контекста. Автоматика и телемеханика, вып. 12, 2020 г., стр. 153-172 / Glazkova A. V. Topical classification of text fragments accounting for their nearest context. Automation and Remote Control, vol. 81, issue 12, pp. 2262-2276.

[18] John M., Jayasudha J.S. Enhancing Performance of Deep Learning Based Text Summarizer. International Journal of Applied Engineering Research, vol. 12, no. 24, 2017, pp. 15986-15993.

[19] Akkasi A., Varoğlu E., Dimililer N. Balanced undersampling: a novel sentence-based undersampling method to improve recognition of named entities in chemical and biomedical text. Applied Intelligence, vol. 48, issue 8, 2018, pp. 1965-1978.

[20] Luo Y., Feng H. et al A novel oversampling method based on SeqGAN for imbalanced text classification. In Proc. of the 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 2891-2894.

[21] Li Y., Guo H. et al. Imbalanced text sentiment classification using universal and domain-specific knowledge. Knowledge-Based Systems, vol. 160, 2018, pp. 1-15.

[22] Chawla N.V., Bowyer K.W.et al. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, vol. 16, 2002, pp. 321-357.

[23] Taha A.Y., Tiun S. et al. Multilabel Over-sampling and Under-sampling with Class Alignment for Imbalanced Multilabel Text Classification. Journal of Information and Communication Technology, vol. 20, issue 3, pp. 423-456.

[24] Gallant S.I. A practical approach for representing context and for performing word sense disambiguation using neural networks. Neural Computation, vol. 3, issue 3, 1991, pp. 293-309.

[25] Huang E.H., Socher R. et al. Improving word representations via global context and multiple word prototypes. In Proc. of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2012, pp. 873-882.
[26] Devlin J., Chang M.W. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv.1810.04805, 2019, 16 p.
[27] Brown T.B., Mann B. et al. Language models are few-shot learners. arXiv.2005.14165, 2020, 75 p.
[28] Cohan A., Dernoncourt F. A discourse-aware attention model for abstractive summarization of long documents. arXiv.1804.05685, 2018, 7 p.
[29] ExplainaBoard – Named Entity Recognition URL: http://explainaboard.nlpedia.ai/leaderboard/task-ner/, accessed 16.05.2022.

Информация об авторах / Information about authors

Федор Владимирович КРАСНОВ – доктор технических наук, эксперт департамента информационных технологий управления. Область научных интересов: интеллектуальная аналитика текстов.

Fedor Vladimirovich KRASNOV – Doctor of Technical Sciences, expert of the Department of Information Technologies of Management. Research interests: intellectual analytics of texts.

Ирина Сергеевна СМАЗНЕВИЧ – бизнес-аналитик, департамент семантических систем. Область научных интересов: применение интеллектуальных алгоритмов в прикладных информационных системах.

Irina Sergeevna SMAZNEVICH – Business Analyst, Department of Semantic Systems. Research interests: application of intelligent algorithms in applied information systems.

Елена Николаевна БАСКАКОВА – ведущий системный аналитик, департамент семантических систем. Область научных интересов: применение интеллектуальных алгоритмов в прикладных информационных системах.

Elena Nikolaevna BASKAKOVA – Leading Systems Analyst, Semantic Systems Department. Research interests: application of intelligent algorithms in applied information systems.