УДК 004.912 + 004.8
DOI 10.25205/1818-7900-2021-19-2-5-16

Метод автоматического извлечения терминов из научных статей на основе слабо контролируемого обучения

Е. П. Бручес, Т. В. Батура

Институт систем информатики им. А. П. Ершова СО РАН
Новосибирск, Россия

Новосибирский государственный университет
Новосибирск, Россия

Аннотация

Описывается метод извлечения научных терминов из текстов на русском языке, основанный на слабо контролируемом обучении (weakly supervised learning). Особенность данного метода заключается в том, что для него не нужны размеченные вручную данные, что является очень актуальным. Для реализации метода мы собрали в полуавтоматическом режиме словарь терминов, затем автоматически разметили тексты научных статей этими терминами. Полученные тексты мы использовали для обучения модели. Затем этой моделью были автоматически размечены другие тексты. Вторая модель была обучена на объединении текстов, размеченных словарем и первой моделью. Результаты показали, что добавление данных, полученных даже автоматической разметкой, улучшает качество извлечения терминов из текстов.

Ключевые слова

извлечение терминов, нейросетевые модели языка, словарный метод, слабо контролируемое обучение, распознавание сущностей, обработка текстов

Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-07-01134

Для цитирования

Бручес Е. П., Батура Т. В. Метод автоматического извлечения терминов из научных статей на основе слабо контролируемого обучения // Вестник НГУ. Серия: Информационные технологии. 2021. Т. 19, № 2. С. 5–16. DOI 10.25205/1818-7900-2021-19-2-5-16

Method for Automatic Term Extraction from Scientific Articles Based on Weak Supervision

E. P. Bruches, T.V. Batura

A. P. Ershov Institute of Informatics Systems SB RAS
Novosibirsk, Russian Federation

Novosibirsk State University
Novosibirsk, Russian Federation

Abstract

We propose a method for scientific terms extraction from the texts in Russian based on weakly supervised learning. This approach doesn't require a large amount of hand-labeled data. To implement this method we collected a list of terms in a semi-automatic way and then annotated texts of scientific articles with these terms. These texts we used to train a model. Then we used predictions of this model on another part of the text collection to extend the train set. The second model was trained on both text collections: annotated with a dictionary and by a second model. Obtained results showed that giving additional data, annotated even in an automatic way, improves the quality of scientific terms extraction.

Keywords

term extraction, neural network language models, dictionary approach, weakly supervised learning, entity recognition, text processing

Acknowledgements

The study was funded by RFBR according to the research project N 19-07-01134

For citation

Bruches E. P., Batura T. V. Method for Automatic Term Extraction from Scientific Articles Based on Weak Supervision. *Vestnik NSU. Series: Information Technologies*, 2021, vol. 19, no. 2, p. 5–16. (in Russ.) DOI 10.25205/1818-7900-2021-19-2-5-16

Введение

В настоящее время всё большую актуальность приобретает автоматическая обработка текстов научных статей, которая включает в себя многие задачи: извлечение терминов, извлечение отношений между терминами, автореферирование и др. Это связано с ростом числа публикуемых работ. Чтение и анализ публикаций в поисках нужной информации становится всё более затратной по времени и ресурсам задачей. В связи с этим в последнее время появляются различные системы автоматического анализа научных статей, например ScholarPhi [1].

В данной статье мы предлагаем метод автоматического извлечения терминов из текстов научных статей на русского языке. Эта задача осложняется тем, что для русского языка не только нет размеченных данных в достаточном количестве, но также довольно проблематично получить неразмеченные тексты статей. Мы разработали метод, не требующий размеченных вручную данных. При этом достигнутые результаты показывают, что данный модуль может быть использован в последующих задачах, требующих информацию о научных терминах.

В данной работе под термином понимается слово или словосочетание, являющееся названием некоторого понятия какой-нибудь области науки, техники, искусства и др. [2]. Общая идея, которая лежит в основе традиционных подходов, состоит в том, что автоматическое извлечение терминов происходит в два этапа: на первом этапе из текстов извлекаются *n*-граммы слов, которые потенциально могут быть терминами, а на втором этапе выполняется классификация, в результате которой принимается решение, является ли данная фраза термином. Алгоритмы, архитектура которых соответствует этой идее, можно также разделить на несколько групп.

Первая группа предполагает использование правил для выделения из текстов фраз, которые являются терминами. Например, в работе [3] предлагается использование словарей и информации о синтаксической структуре предложения для извлечения многословных терминов.

Другую группу составляют методы, в основе которых лежит использование алгоритмов машинного обучения с вручную извлеченными признаками. Так, в работе [4] описывается алгоритм извлечения как однословных, так и многословных терминов: на первом этапе из текстов извлекаются *n*-граммы слов, которые потенциально могут быть терминами, а затем на основании различных входных признаков алгоритм определяет, является ли *n*-грамма термином. В статье [5] авторы используют несколько групп признаков для извлечения терминов: лингвистические (части речи, главное слово фразы, количество имен существительных во фразе и др.), статистические (длина фразы, TF, IDF, TF-IDF и др.) и гибридные признаки (например, частота встречаемости фразы в корпусах обычных и научных текстов). Также было исследовано применение алгоритма PageRank для более точной классификации [6]. В работе [7] предлагается использовать признаки, основанные на информации из Википедии.

В третью группу входят методы глубокого обучения. В работе [8] исследуется проблема отсутствия достаточного количества данных. Для этого авторы на ограниченном количестве

данных обучают две модели (CNN и LSTM), которые на вход принимают векторные представления слов фразы, а на выходе определяют, является данная фраза термином или нет. Затем этими моделями размечается новая порция данных, которая добавляется в обучающую выборку, и процесс обучения повторяется еще раз.

В работе [9] предлагается архитектура, состоящая из этапов, отличных от тех, что были описаны: на первом шаге классификатор определяет, содержит ли входное предложение термины; если содержит, то на втором этапе происходит непосредственно нахождение терминов в предложении.

Другая общая идея – рассматривать задачу извлечения терминов как задачу сопоставления последовательностей входных токенов с последовательностями меток из заранее определенного множества (sequence labelling task), т. е. для каждого токена в тексте требуется определить его класс (является он термином или нет). Таким образом, решение задачи осуществляется в один этап. Как правило, при таком подходе используется разметка сущностей в формате BIO (BILOU и др.). Большим преимуществом данного подхода является то, что во внимание принимается контекст (как семантический, так и синтаксический) употребления конкретной фразы, что составляет один из ключевых признаков для нахождения терминов в тексте. Так, в работе [10] исследуются различные архитектуры и векторные представления слов при решении задачи sequence labelling.

Еще одна идея, которая отличается от описанных выше, состоит в использовании методов тематического моделирования для извлечения терминов. В статье [11] описывается попытка применения различных методов тематического моделирования для улучшения нахождения однословных терминов: невероятностные (разные методы кластеризации – K-means, NFM и др.) и вероятностные (в качестве метода такой группы был выбран алгоритм LDA).

Алгоритм извлечения терминов

Ввиду отсутствия достаточного количества размеченных данных для задачи извлечения терминов для русского языка мы приняли решение использовать подход псевдоразметки (pseudo-labelling). Он заключается в том, чтобы обучить модель на небольшом количестве размеченных данных, а затем разметить полученной моделью некоторое количество новых текстов, добавить их к обучающему множеству и обучить вторую модель.

Таким образом, алгоритм получения модели для извлечения терминов состоит из следующих шагов:

1) получить размеченный корпус для первой итерации обучения модели с помощью словарного подхода;

2) обучить модель на полученном корпусе из п. 1;

3) разметить новые тексты и тексты из п. 1 моделью, полученной в результате выполнения п. 2, и словарным подходом;

4) обучить модель на полученном корпусе текстов из п. 3.

Схема алгоритма представлена на рис. 1.

Рассмотрим каждый из шагов более детально.

Получение размеченного корпуса для первой итерации обучения модели

Идея этого подхода состоит в использовании заранее составленного словаря терминов. Словарь терминов набирался в полуавтоматическом режиме двумя способами.

1. Из текстов научных статей были автоматически извлечены 2-, 3- и 4-граммы слов, отсортированные по значению tf-idf, затем из них вручную были выбраны те фразы, которые потенциально могут быть терминами.

2. Из заголовков статей из Википедии, входящих в подграф категории «Наука», были вручную выбраны те слова и фразы, которые потенциально могут быть терминами.

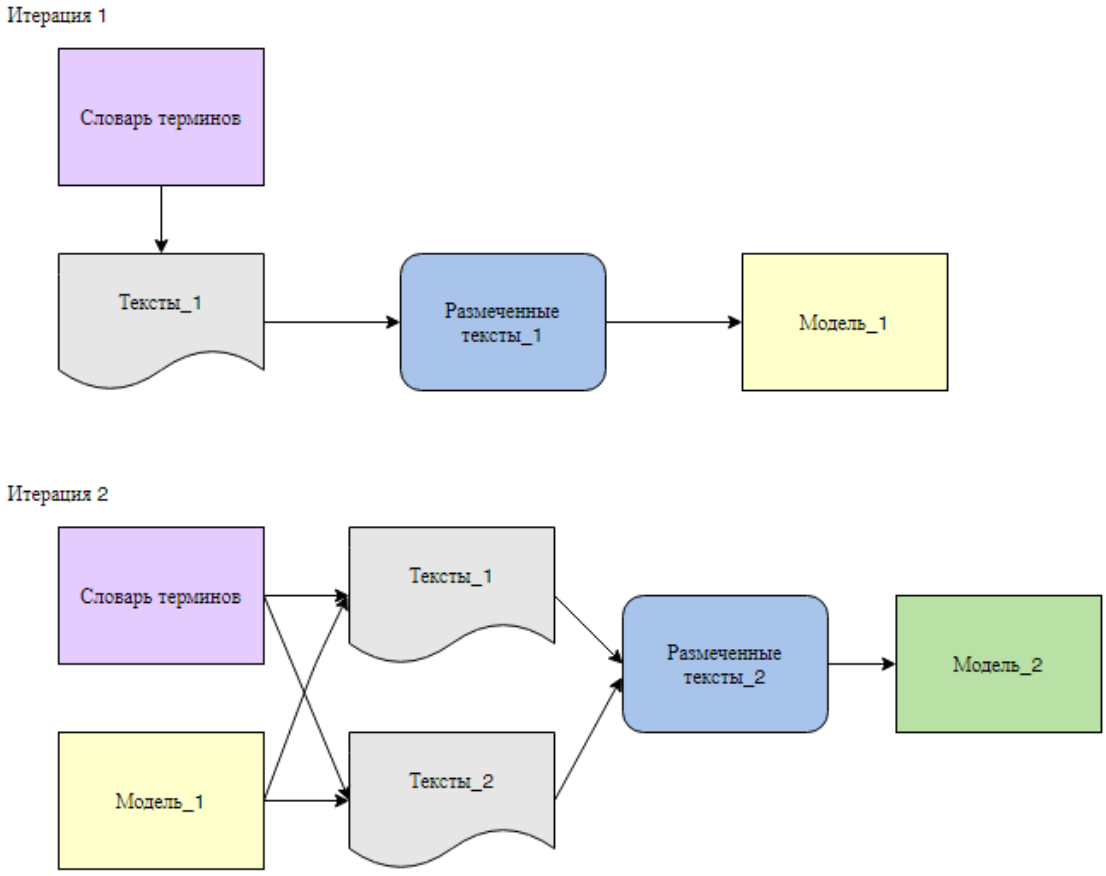

Рис. 1. Схема алгоритма
Fig. 1. Algorithm

Таким образом был получен словарь из 17 252 терминов:

- 1-граммы (6 509 терминов): *этнолингвистика*, *равносоставленность*, *пластида* и др.;
- 2-граммы (6 785 терминов): *нокдаун гена*, *математическое ожидание* и др.
- 3-граммы (2 322 терминов): *уравнение переноса излучений*, *метод анализа иерархий* и др.;
- 4-граммы (1 098 терминов): *теорема Бертрана о выборах*, *стабилизированный метод бисопряжённых градиентов* и др.;
- 5-граммы (348 терминов): *теория реакционной способности химических соединений* и др.;
- 6-граммы (120 терминов): *теорема Римана об условно сходящихся рядах*, *задача о назначении минимального количества исполнителей* и др.;
- 7-граммы (40 терминов): *теорема о свойстве Дарбу для непрерывной функции* и др.;
- 8-граммы (22 термина): *теорема Пуанкаре о разложении интегралов по малому параметру* и др.;
- 9-граммы (7 терминов): *теорема Стоуна о группах унитарных операторов в гильбертовом пространстве* и др.;
- 12-граммы (1 термин): *автоматический выключатель*, *управляемый дифференциальным током*, *со встроенной защитой от сверхтока*.

Основная сложность составления словаря терминов заключается в том, что без контекста сложно понять, является фраза термином или нет. Более того, в разных контекстах одна и та же фраза может быть как термином, так и нет; например, *модель*, *текст*, *язык* и др.

Таким образом, было размечено 1 118 текстов научных статей из журнала «Программные системы и продукты», которые использовались в качестве обучающего множества для модели в первой итерации (Bert-iter_1). Полученный словарь терминов находится в открытом доступе [1].

Следует отметить, что этап создания словаря может быть заменен на этап пополнения уже имеющегося словаря терминов выбранной предметной области (при наличии подходящего). Для экспериментов мы выбрали область компьютерных технологий как пример быстро меняющейся отрасли, в которой в настоящий момент не вся терминология является устоявшейся: одни термины появляются, другие быстро устаревают. В таких условиях довольно сложно составлять словари терминов и поддерживать их в актуальном состоянии.

Получение размеченного корпуса для второй итерации обучения модели

На следующем шаге мы взяли 808 аннотаций научных статей из журналов «Cloud of science», «Программные системы и вычислительные методы», «Информационно-управляющие системы» и разметили их комбинированной моделью. Данная модель (Bert-iter_2) представляет собой комбинацию словарного метода, с помощью которого были размечены тексты в предыдущем пункте, и модели, полученной после первой итерации обучения.

Обучающим множеством для модели второй итерации стало объединение текстов первой итерации и новых размеченных текстов.

Основная гипотеза при использовании предсказаний модели для разметки текстов состоит в том, что обобщающая способность модели позволяет не только выделять термины, которые уже содержатся в словаре, но находить новые паттерны и, соответственно, новые слова и фразы, которые являются терминами.

Описание модели

В обеих итерациях мы использовали одну и ту же архитектуру нейронной сети. Векторные представления слов для входных текстов мы получали, используя предобученную модель bert-base-multilingual-cased [2]. Затем идут слой двунаправленной LSTM и два полносвязных слоя. Архитектура модели показана на рис. 2.

На вход модели подается токенизированный текст (входные тексты предварительно никак не обрабатываются). Выход модели представляет собой последовательность предсказанных классов для соответствующих токенов. В данной задаче используется набор из трех классов: «B-TERM» (обозначает первый токен в термине, *beginning*)), «I-TERM» (обозначает не первый токен в термине, *inside*), «O» (обозначает токен, который не входит в состав термина, *outside*).

Для улучшения качества извлечения терминов, мы также написали валидацию эвристиками. Ниже приводится описание используемых эвристик.

Описание эвристик

Эвристика 1: если токен (1) распознан как термин и является именем существительным или именем прилагательным, и следующий токен (2) распознан как не термин и имеет форму родительного падежа, то токену (2) присваивается тэг «I-TERM». Примеры последовательностей: *методы сжатия (1) данных (2)*, *реляционных баз (1) данных (2)*, *системы (1) проверки (2) заданий* и др.

[1] https://github.com/iis-research-team/ner-rc-russian/tree/master/dict_extractor
[2] https://huggingface.co/transformers/pretrained_models.html

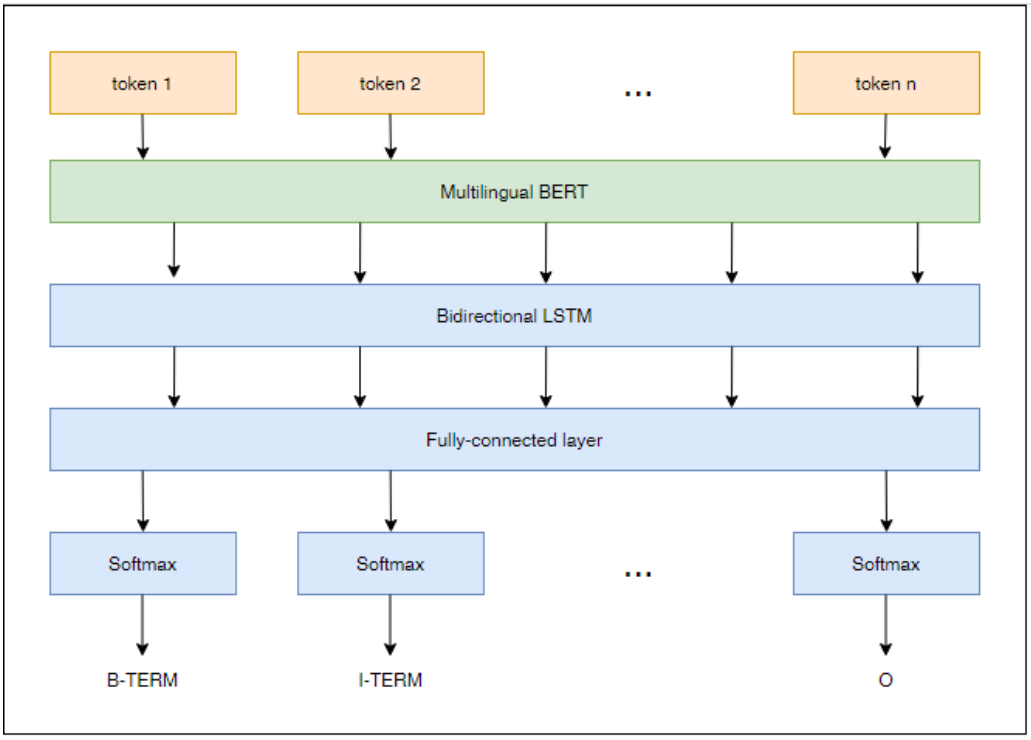

Рис. 2. Архитектура нейронной сети
Fig. 2. Neural network architecture

Эвристика 2: если токен (1) распознан как «B-TERM» и является именем прилагательным, и токен (2) распознан как «B-TERM» и является именем существительным, то меняем тэг токена (2) на «I-TERM». Примеры последовательностей: *в учебном (1) процессе (2)*, *нечёткая (1) модель (2)*, *физические (1) величины (2)* и др.

Эвристика 3: если последний токен (1) в цепочке токенов, образующих термин, имеет часть речи имя прилагательное, а следующий за ним токен имеет часть речи имя существительное (2), то токену (2) присваивается тэг «I-TERM». Примеры последовательностей: *возможности мобильных (1) устройств (2)* и др.

Эвристика 4: если токен (1) входит в состав термина, а следующий за ним токен состоит только из латинских символов, то включаем его в состав термина. Пример последовательностей: *в пакете (1) MOST (2)*, *по базе данных (1) Cistrome (2)* и др.

Также мы в явном виде запрещаем помечать тэгами терминов следующие классы токенов:

1) знаки пунктуации: «.», «,», «:», «;»;

2) предлоги и союзы, если они имеют тэг «B-TERM» (мы допускаем появление предлогов и союзов в составе термина, но не допускаем того, что данные части речи начинают термин);

3) однозначные глаголы и деепричастия (под однозначными подразумевается, что токены не имеют морфологической омонимии).

Анализ результатов

Полученные на каждом шаге модели мы оценивали на корпусе RuSERRC [12], который не использовался в процессе обучения.

Для оценки качества моделей мы использовали стандартные метрики классификации: точность, полноту и F-меру. Точность – это доля верно извлеченных терминов относительно

всех терминов, которые извлекла модель. Полнота – это доля верно найденных моделью терминов относительно всех терминов в тестовой выборке. F-мера представляет собой гармоническое среднее между точностью и полнотой (в табл. 1 приведена взвешенная F-мера).

При этом мы рассматривали несколько вариантов того, что имеется в виду под словом «термин».

В первом случае под термином понимается вся последовательность токенов, входящая в состав термина. Если хотя бы один токен распознан неверно, то считается, что весь термин распознан неверно. В табл. 1 такая метрика указана как «Полное совпадение».

Во втором случае под термином понимается токен, тэг которого принадлежит множеству {“B-TERM”, “I-TERM”}. Ввиду большой неоднозначности определения границ терминов, которая присутствовала также при разметке корпуса асессорами, данная метрика видится нам релевантной – она показывает, насколько хорошо модель способна распознать фразы, которые могут быть терминами, без указания точных границ. В табл. 1 такая метрика указана как «Частичное совпадение».

Для сравнения мы взяли алгоритм извлечения ключевых слов RAKE, применение которого к этому корпусу описано в статье [13]. Rapid automatic keyword extraction (RAKE) – алгоритм, предназначенный для автоматического извлечения ключевых слов [14]. Сначала применяется список стоп-слов и разделителей для выделения многословных терминов, после чего используется статистическая информация: для каждого слова из ключевых фраз-кандидатов оценивается частота, с которой оно встречалось, и количество связей между этим словом и остальными. На основании этих двух величин вычисляется вес ключевой фразы, и все фразы сортируются по весам, а наиболее вероятные ключевые фразы получают максимальный вес. Оригинальный алгоритм в качестве ключевых слов выделяет также и глагольные формы, но так как в данной задаче рассматриваются только именные группы, то в статье описан оптимизированный алгоритм RAKE, в котором на этапе предобработки текста удаляются все глагольные формы. Полученные метрики приведены в табл. 1.

Таблица 1

Метрики извлечения терминов

Table 1

Term Extraction Metrics

Метод	Полное совпадение			Частичное совпадение		
	Точность	Полнота	F-мера	Точность	Полнота	F-мера
Bert-iter_1	0,22	0,19	0,20	0,76	0,71	0,68
Bert-iter_1 + эвристики	0,39	0,28	0,33	0,76	0,75	0,74
Bert-iter_2	0,30	0,25	0,28	0,77	0,75	0,74
Bert-iter_2 + эвристики	0,40	0,29	0,34	0,77	0,77	0,76
Bert-iter_2 + эвристики + словарь	0,39	0,31	0,35	**0,78**	**0,78**	**0,77**
Rake	0,36	0,28	0,32	0,62	0,63	0,63
Оптимизированный Rake	**0,44**	**0,35**	**0,39**	0,65	0,57	0,61

Относительно низкие метрики во многом связаны с различием разметки обучающего множества и золотого стандарта. В силу того что обучающее множество было получено с помощью автоматической разметки терминов из словаря, последовательность токенов, являющаяся термином, не претерпевала никаких изменений. В реальных текстах термин может быть «разорван», содержать синонимы, сокращения, пунктуационные знаки или вовсе быть неполным.

Если проанализировать метрики частичного совпадения токенов, то видно, что модель способна находить места в тексте, в которых может быть термин, без точного определения границ термина. Учитывая, что задача определения границ термина является достаточно сложной даже для человека (что, например, показывает метрика согласованности асессоров при разметке сущностей в корпусе), то полученные метрики кажутся нам достаточными для использования данного подхода при решении других задач (например, задачи связывания именованных сущностей или классификации отношений между сущностями).

Применение модели к текстам другой предметной области

Для этого эксперимента мы использовали размеченные тексты из корпуса RuREBus [15], который содержит программы стратегического развития, предоставленные Минэкономразвития РФ. Так как мы работаем с текстами из области математики и информационных технологий, то область экономики как раз подходит для проверки способности модели к обобщению. Более того, жанры текстов также различаются: модель была обучена на текстах научных статей, в то время как в корпусе RuREBus содержатся тексты различных экономических документов. В разметке этого корпуса используется 8 типов именованных сущностей:

1) метрика – индикатор или объект, на основании которого производится сравнение (например, *уровень рождаемости*, *экономический рост*);

2) экономика – экономическая сущность или объект инфраструктуры (например, *ПАО Газпром*, *библиотечные и музейные фонды*);

3) институт – институты, структуры и организации (например, *Центр занятости молодёжи*, *системы дорог*);

4) бинарный – бинарная характеристика или единичное действие (например, *модернизация*, *функционирует*);

5) сравнение – сравнительная характеристика (например, *рост*, *негативная динамика*);

6) качество – качественная характеристика (например, *эффективный*, *безопасный*);

7) социальный – социальный объект (например, *научный и образовательный потенциал*, *досуг*);

8) деятельность – деятельность, события (например, *реставрационные работы*, *ярмарка выходного дня*).

Из приведенных примеров видно, что под сущностями авторы RuREBus понимают не только именные группы, но и глаголы и глагольные группы, и отдельные прилагательные. Такой принцип выделения именованных сущностей расходится с тем, которому мы следуем в данной работе, – под именованными сущностями мы подразумеваем только существительные и именные группы. Поэтому нам кажется некорректным непосредственное сравнение метрик. Тем не менее мы можем проанализировать результаты и сделать некоторые выводы.

В табл. 2 представлены метрики полного и частичного совпадений, полученные нашей моделью на корпусе RuREBus. Очевидно, что модель плохо находит границы терминов, но способна находить слова и фразы, которые входят в состав термина.

Таблица 2

Метрики на корпусе RuREBus

Table 2

Metrics for RuREBus

Метод	Полное совпадение			Частичное совпадение		
	Точность	Полнота	F-мера	Точность	Полнота	F-мера
Bert-iter_2 + эвристики + словарь	0,17	0,13	0,15	0,68	0,36	0,47

Таблица 3

Примеры выделения терминов из текстов RuREBus

Table 3

Examples of terms extraction from RuREBus texts

№ п/п	Текст
1	*Повышение **[доступности транспортных услуг]** для **[населения]**.*
2	*Организация **[временного трудоустройства отдельных категорий безработных граждан]**.*
3	***[Право] [граждан]** на благоприятную **[среду жизнедеятельности]** закреплено в основном **[законе государства] - [Конституции Российской Федерации]**.*
4	*Контроль за исполнением **[постановления]** возложить на **[первого заместителя руководителя администрации]** МР "Усть-Вымский" Карпову А.Д.*
5	***[Оценка эффективности рисков] - [риски]** низкие.*
6	***[Деятельность учреждений культуры]** и искусства является одной из важнейших составляющих современной культурной жизни.*
7	***[ПОСТАНОВЛЕНИЕ]** об утверждении муниципальной программы "**[Поддержка сельскохозяйственных товаропроизводителей]** и **[создание условий]** для **[развития сферы заготовки]** и **[переработки дикорастущего сырья Верхнекетского района]** на 2016-2021 годы".*
8	***[Паспорт] [муниципальной программы]** городского округа Кашира "**[Спорт городского округа]** Кашира" на 2017-2021 годы.*
9	*Приложение к **[постановлению администрации муниципального имущества]***
10	*За этот период было установлено и заменено 366 дорожных **[знаков]** и 43 сигнальных столбика на **[железнодорожных переездах]**.*

В табл. 3 приведены примеры полученной нами разметки предложений из корпуса RuREBus с помощью предложенного подхода – в квадратных скобках заключены выделенные сущности. Видно, что в тексте автоматически выделяются последовательности токенов, которые являются сущностями в данном контексте. Таким образом, можно заключить, что данный метод потенциально может применяться к текстам из различных предметных областей и давать приемлемые результаты в пределах заданного языка (в нашем случае русского) без дополнительных затрат.

Заключение

В данной работе описан метод автоматического извлечения терминов из научных текстов на русском языке. Метод основан на слабо контролируемом обучении и позволяет решать задачу без размеченных вручную данных, что является особенно актуальным на сегодняшний день. Полученные результаты показали эффективность использования дополнительных данных, размеченных даже автоматически. Кроме того, анализ применения данной модели к текстам другой области знаний показал, что модель способна к обобщению.

Список литературы

1. **Head A.**, **Lo K.**, **Kang D.**, **Fok R.**, **Skjonsberg S.**, **Weld D. S.**, **Hearst M. A.** Augmenting Scientific Papers with Just-in-Time, Position-Sensitive Definitions of Terms and Symbols. ArXiv: 2009.14237. 2021.

2. **Лопатин В. В.**, **Лопатина Л. Е.** Русский толковый словарь: около 35 000 слов. М.: Русский язык, 1997. 832 с.
3. **Stanković R.**, **Krstev C.**, **Obradović I.**, **Lazić B.**, **Trtovac A.** Rule-based Automatic Multi-word Term Extraction and Lemmatization. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16). 2016, p. 507–514.
4. **Yuan Y.**, **Gao J.**, **Zhang Y.** Supervised Learning for Robust Term Extraction. In: Proceedings of 2017 International Conference on Asian Language Processing (IALP). 2017, p. 302–305. DOI 10.1109/IALP.2017.8300603
5. **Conrado M.**, **Pardo T.**, **Rezende S. O.** A Machine Learning Approach to Automatic Term Extraction using a Rich Feature Set. In: Proceedings of the NAACL HLT 2013 Student Research Workshop. Atlanta, Georgia, 2013, p. 16–23.
6. **Zhang Z.**, **Gao J.**, **Ciravegna F.** SemRe-Rank: Improving Automatic Term Extraction by Incorporating Semantic Relatedness with Personalised PageRank. *ACM Transactions on Knowledge Discovery from Data* (*TKDD*), 2018, vol. 12, no. 5, p. 1–41.
7. **Bilu Y.**, **Gretz Sh.**, **Cohen E.**, **Slonim N.** What if we had no Wikipedia? Domain-independent Term Extraction from a Large News Corpus. arXiv: 2009.08240. 2020.
8. **Wang R.**, **Liu W.**, **McDonald C.** Featureless Domain-Specific Term Extraction with Minimal Labelled Data. In: Proceedings of Australasian Language Technology Association Workshop, 2016, p. 103−112.
9. **Hossari M.**, **Dev S.**, **Kelleher J. D.** TEST: A Terminology Extraction System for Technology Related Terms. In: Proceedings of the 2019 11th International Conference on Computer and Automation Engineering, 2019, p. 78–81. DOI 10.1145/3313991.3314006
10. **Kucza M.**, **Niehues J.**, **Zenkel T.**, **Waibel A.**, **Stüker S.** Term Extraction via Neural Sequence Labeling a Comparative Evaluation of Strategies Using Recurrent Neural Networks. In: Proceedings of Interspeech 2018. 2018. p. 2072-2076.
11. **Bolshakova E.**, **Loukachevitch N.**, **Nokel M.** Topic Models Can Improve Domain Term Extraction. In:European Conference on Information Retrieval (ECIR 2013). Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2013, vol. 7814, p. 684–687.
12. **Bruches E.**, **Pauls A.**, **Batura T.**, **Isachenko V.** Entity Recognition and Relation Extraction from Scientific and Technical Texts in Russian. In: Science and Artificial Intelligence conference, 2020, p. 41–45. DOI 10.1109/S.A.I.ence50533.2020.9303196
13. **Бручес Е. П.**, **Паульс А. Е.**, **Батура Т. В.**, **Исаченко В. В.**, **Щербатов Д. Р.** Семантический анализ научных текстов: опыт создания корпуса и построения языковых моделей // Программные продукты и системы, 2020, 18 с.
14. **Rose S.**, **Engel D.**, **Cramer N.**, **Cowley W.** Automatic Keyword Extraction from Individual Documents. *Text Mining: Theory and Applications*, 2010, vol. 1, p. 1–20. DOI 10.1002/9780470689646.ch1
15. **Ivanin V.**, **Artemova E.**, **Batura T.**, **Ivanov V.**, **Sarkisyan V.**, **Tutubalina E.**, **Smurov I.** RUREBUS-2020 Shared Task: Russian Relation Extraction for Business. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog", 2020, p. 416–431. DOI 10.28995/2075-7182-2020-19-416-431

References

1. **Head A.**, **Lo K.**, **Kang D.**, **Fok R.**, **Skjonsberg S.**, **Weld D. S.**, **Hearst M. A.** Augmenting Scientific Papers with Just-in-Time, Position-Sensitive Definitions of Terms and Symbols. ArXiv: 2009.14237. 2021.
2. **Lopatin V. V.**, **Lopatina L. E.** Russian Explanatory Dictionary. Moscow, 1997, 832 p. (in Russ.)

3. **Stanković R.**, **Krstev C.**, **Obradović I.**, **Lazić B.**, **Trtovac A.** Rule-based Automatic Multi-word Term Extraction and Lemmatization. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16). 2016, p. 507–514.
4. **Yuan Y.**, **Gao J.**, **Zhang Y.** Supervised Learning for Robust Term Extraction. In: Proceedings of 2017 International Conference on Asian Language Processing (IALP). 2017, p. 302–305. DOI 10.1109/IALP.2017.8300603
5. **Conrado M.**, **Pardo T.**, **Rezende S. O.** A Machine Learning Approach to Automatic Term Extraction using a Rich Feature Set. In: Proceedings of the NAACL HLT 2013 Student Research Workshop. Atlanta, Georgia, 2013, p. 16–23.
6. **Zhang Z.**, **Gao J.**, **Ciravegna F.** SemRe-Rank: Improving Automatic Term Extraction by Incorporating Semantic Relatedness with Personalised PageRank. *ACM Transactions on Knowledge Discovery from Data* (*TKDD*), 2018, vol. 12, no. 5, p. 1–41.
7. **Bilu Y.**, **Gretz Sh.**, **Cohen E.**, **Slonim N.** What if we had no Wikipedia? Domain-independent Term Extraction from a Large News Corpus. arXiv: 2009.08240. 2020.
8. **Wang R.**, **Liu W.**, **McDonald C.** Featureless Domain-Specific Term Extraction with Minimal Labelled Data. In: Proceedings of Australasian Language Technology Association Workshop, 2016, p. 103−112.
9. **Hossari M.**, **Dev S.**, **Kelleher J. D.** TEST: A Terminology Extraction System for Technology Related Terms. In: Proceedings of the 2019 11th International Conference on Computer and Automation Engineering, 2019, p. 78–81. DOI 10.1145/3313991.3314006
10. **Kucza M.**, **Niehues J.**, **Zenkel T.**, **Waibel A.**, **Stüker S.** Term Extraction via Neural Sequence Labeling a Comparative Evaluation of Strategies Using Recurrent Neural Networks. In: Proceedings of Interspeech 2018. 2018. p. 2072-2076.
11. **Bolshakova E.**, **Loukachevitch N.**, **Nokel M.** Topic Models Can Improve Domain Term Extraction. In:European Conference on Information Retrieval (ECIR 2013). Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2013, vol. 7814, p. 684–687.
12. **Bruches E.**, **Pauls A.**, **Batura T.**, **Isachenko V.** Entity Recognition and Relation Extraction from Scientific and Technical Texts in Russian. In: Science and Artificial Intelligence conference, 2020, p. 41–45. DOI 10.1109/S.A.I.ence50533.2020.9303196
13. **Bruches E.**, **Pauls A.**, **Batura T.**, **Isachenko V.**, **Shcherbatov D.** Semantic Analysis of Scientific Texts: Experience in Creating a Corpus and Building Language Models. *Software & Systems*, 2020, 18 p. (in Russ.)
14. **Rose S.**, **Engel D.**, **Cramer N.**, **Cowley W.** Automatic Keyword Extraction from Individual Documents. *Text Mining: Theory and Applications*, 2010, vol. 1, p. 1–20. DOI 10.1002/9780470689646.ch1
15. **Ivanin V.**, **Artemova E.**, **Batura T.**, **Ivanov V.**, **Sarkisyan V.**, **Tutubalina E.**, **Smurov I.** RUREBUS-2020 Shared Task: Russian Relation Extraction for Business. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog", 2020, p. 416–431. DOI 10.28995/2075-7182-2020-19-416-431

Материал поступил в редколлегию
Received
15.02.2021

Сведения об авторах

Батура Татьяна Викторовна, кандидат физико-математических наук, старший научный сотрудник, Институт систем информатики им. А. П. Ершова СО РАН (Новосибирск, Россия); доцент, Новосибирский государственный университет (Новосибирск, Россия)

tatiana.v.batura@gmail.com
ORCID 0000-0003-4333-7888

Бручес Елена Павловна, аспирант, Институт систем информатики им. А. П. Ершова СО РАН (Новосибирск, Россия); ассистент, Новосибирский государственный университет (Новосибирск, Россия)

bruches@bk.ru

Information about the Authors

Tatiana V. Batura, PhD in Physics and Mathematics, Senior Researcher, A. P. Ershov Institute of Informatics Systems SB RAS (Novosibirsk, Russian Federation); Associate Professor, Novosibirsk State University (Novosibirsk, Russian Federation)

tatiana.v.batura@gmail.com
ORCID 0000-0003-4333-7888

Elena P. Bruches, PhD student, A. P. Ershov Institute of Informatics Systems SB RAS (Novosibirsk, Russian Federation); Assistant, Novosibirsk State University (Novosibirsk, Russian Federation)

bruches@bk.ru