УДК 004.912+004.8
DOI: 10.15827/0236-235X.140.689-697
Дата подачи статьи: 02.07.22, после доработки: 24.08.22
2022. Т. 35. № 4. С. 689–697

Извлечение терминов из текстов научных статей

Я.Ю. Дементьева [1], *студент, y.dementeva@g.nsu.ru*
Е.П. Бручес [1, 2], *младший научный сотрудник, ст. преподаватель, bruches@bk.ru*
Т.В. Батура [2], *к.ф.-м.н., доцент, старший научный сотрудник, tatiana.v.batura@gmail.com*

[1] *Новосибирский государственный университет, г. Новосибирск, 630090, Россия*
[2] *Институт систем информатики им. А.П. Ершова СО РАН, г. Новосибирск, 630090, Россия*

Актуальность задачи извлечения терминов из текстов научных статей обусловлена необходимостью автоматического аннотирования и выделения ключевых слов в постоянно увеличивающемся потоке научно-технических документов. В данной статье исследуется влияние различных языковых моделей на качество извлечения научных терминов из текстов на русском языке.

Для сравнения были взяты две модели: mBERT, предварительно обученная на текстах разных языков, и ruBERT, предобученная только на русскоязычных данных. В ходе работы были подготовлены два обучающих множества – размеченное вручную и размеченное автоматически. С применением этих обучающих множеств были проведены дообучение и дальнейшее сравнение показателей работы двух моделей. Изучено влияние выбора языковой модели на численные показатели качества извлечения терминологии, содержащейся в текстах научных статей.

На основе полученных результатов был модернизирован алгоритм извлечения терминологии из текстов инструментом Terminator.

В рамках задачи извлечения терминологии из текстов научных статей на русском языке наиболее применимой можно считать модель ruBERT, давшую лучшие показатели в ансамбле со словарем и эвристиками. Кроме того, можно констатировать разницу результатов моделей на полном и на частичном совпадении, обусловленную описанной в статье проблемой определения границ терминов в текстах.

Проведенное исследование позволило также сделать вывод о влиянии качества разметки обучающего набора на качество извлечения терминологии.

Ключевые слова*: извлечение терминологии, NLP, машинное обучение, языковая модель, mBERT, ruBERT, словарь терминов.*

Поток текстовой информации постоянно растет, и многообразие вариантов ее представления на естественном языке вынуждает расширять инструментарий для извлечения. Извлечение терминов (иногда также называемое распознаванием терминов) является подзадачей извлечения информации из текстов, цель которого заключается в автоматическом извлечении релевантных слов и словосочетаний из заданного корпуса текстов [1].

Несмотря на кажущееся многообразие существующих решений, задача извлечения терминов из текстов на естественном языке продолжает оставаться актуальной и требует привлечения новых специалистов. При ее решении могут возникнуть некоторые проблемы.

1. Определение границ составных терминов. Например, рассмотрим составной термин «*многолучевой канал связи с замираниями*» и вариативность определения его границ в тексте: [типичные [многолучевые [каналы связи]] с замираниями].

В данном примере квадратными скобками обозначены возможные границы термина.

2. Распознавание лексической единицы как части составного термина или как свободной лексической единицы.

В тексте могут встретиться такие составные словосочетания, как, например, «*технология конструирования логических моделей*». В таком случае встает вопрос о том, воспринимать ли данное словосочетание как единый термин или стоит разделить его на отдельные: [[[технологии] [конструирования]] [[логических моделей]]].

Возможны следующие варианты:

– «*технологии*» + «*конструирования*» + «*логических моделей*»;

– «*технологии конструирования*» + «*логических моделей*»;

– «*технологии конструирования логических моделей*».

3. Определение лексической единицы как термина в зависимости от контекста и тема-

тики текста, в котором данная лексическая единица употребляется.

Например, в тексте математической тематики слово «*форма*» может быть рассмотрено как часть составного термина «*логическая форма*», тогда как в технических текстах это слово не стоит рассматривать как термин.

Существует множество подходов к решению данных проблем. За последние 50 лет автоматическое извлечение терминологии из научных текстов сформировалось как отдельная отрасль разработки алгоритмов и ПО в рамках компьютерной лингвистики, и в ней были получены существенные достижения. Однако эти методы не дают абсолютных результатов и нуждаются в дальнейшей разработке.

Одним из наиболее актуальных на сегодняшний день подходов к обработке текстов, и к извлечению терминологии в частности, можно считать использование машинного обучения. В данном подходе ключевым вопросом является выбор наиболее подходящей языковой модели, которая после дальнейшего обучения сможет обеспечить лучшее качество извлечения терминов из текстов научных статей.

Цель данной работы – выявление языковой модели для решения задачи извлечения терминов, содержащихся в текстах научных статей на русском языке, для дальнейшей интеграции модели в инструмент, разработанный в Институте систем информатики им. А.П. Ершова СО РАН (https://github.com/iis-research-team/terminator).

Обзор существующих методов

У существующих систем автоматического извлечения терминов можно выделить ряд общих характеристик. Так, на вход системе обычно подается корпус текстов, на выходе система формирует списки терминов-кандидатов, которые подлежат дальнейшей проверке. Рассмотрим различные методы и алгоритмы для извлечения терминов из текста, применяемые (отдельно и в комбинациях) на практике.

Извлечение терминов с опорой на статистические показатели. Вероятностный метод извлечения терминов – это метод, основанный на данных о частотности словосочетания и о совместной встречаемости словосочетаний (возможно использование морфологических шаблонов-фильтров). То есть такие системы работают за счет выделения двух или более лексических единиц, частота совместной встречаемости которых больше некоторого заданного уровня.

Данный метод применяется для извлечения как терминологии в целом [2], так и только ключевых слов для документа [3].

У систем, основанных на таком подходе, есть слабое место: при увеличении длины термина падает частота его встречаемости, в специализированном корпусе ограниченного объема термин может встречаться один-два раза. Можно устанавливать низкий порог встречаемости сочетания слов для включения его в контрольную группу, но тогда метод будет характеризоваться высокой полнотой и, как следствие, низкой точностью выделения терминов.

Извлечение терминов на основе правил. Системы, в основе которых лежит данный метод, функционируют за счет подготовленного исследователями подробного свода правил. В зависимости от правил их можно разделить на три типа:

– системы, в которых с помощью регулярных выражений и конечных автоматов задается выделение повторяющихся структур терминологических сочетаний;

– системы, использующие для автоматической разметки словари;

– системы, использующие общеязыковую информацию (например, о синтаксической структуре предложения, о частеречной принадлежности входящих в него слов), которые выделяют базовые языковые структуры.

В статье [4] рассмотрен метод, опирающийся на информацию о синтаксической структуре предложения и частеречной принадлежности входящих в него слов. Информация о синтаксической структуре хранится в виде шаблона предложения (sentence pattern), а частеречная разметка – в виде последовательности указателей на ту или иную часть речи (POS sequence of the sentence). Авторы утверждают, что такой метод может извлекать терминологию, не требуя корпуса размеченных данных (для холодного старта требуется только словарь шаблонов предложений).

Несколько иначе применены методы на основе правил авторами статьи [5], представившими двухэтапный подход к извлечению тематической информации, относящейся к функциям некоторого ПО, из руководства пользователя. На первом этапе полуавтоматически извлекается терминология предметной области на основе лингвистических шаблонов, а затем применяются методы, опирающиеся на предварительно извлеченную терминологию предметной области и на языковую информацию о структуре предложений.

Метод на основе правил часто применяется с опорой на онтологии и словари. В статье [6] описан метод извлечения многокомпонентных терминов, который опирается на обширные лексические ресурсы в виде электронных словарей и преобразователей с конечным числом состояний для моделирования различных синтаксических структур терминов. Та же технология используется для лемматизации извлеченных многокомпонентных терминов. Извлеченные и лемматизированные многословные термины фильтруются, чтобы отклонить ложно предложенные леммы, а затем ранжируются путем введения показателей, сочетающих лингвистическую и статистическую информацию (C-Value, T-Score, LLR и Keyness).

Системы извлечения терминологии на основе правил дают хорошую точность, но в большинстве случаев только для конкретного языка в силу особенностей его грамматики.

Извлечение терминологии с помощью методов машинного обучения. В последнее время широкое признание получили методы, в основе которых лежит использование алгоритмов машинного обучения. Как правило, они подразумевают два этапа: извлечение цепочек слов, которые потенциально могут быть терминами, и дальнейшее определение термина и уточнение его границ.

В общем случае тексты делятся на значимые интервалы, часто совпадающие с отдельными словами и называемые токенами. Каждому токену соответствует вектор. С множеством таких векторов далее работает модель. В зависимости от специфики задачи при вычислениях весов модели делается упор на предсказание либо токена в контексте, либо контекста для токена. Предварительно обученная на объемном корпусе текстов модель выделяет термины-кандидаты, которые далее могут быть верифицированы. Для определения, является ли последовательность слов термином, могут быть использованы разные признаки: общелингвистическая информация (частеречная принадлежность слов, главное слово фразы, количество имен существительных во фразе и др.) [7], статистические (длина фразы, TF, IDF, TF-IDF или частота встречаемости фразы в корпусе научных текстов, как предложено в [8]) и гибридные признаки [9].

Применяя методы машинного обучения для решения задачи извлечения терминологии из текстов, можно их расширять и совершенствовать. В статье [10] предложено использовать сглаживание границ интервалов при векторизации. Векторизация в span-based модели проводится не для отдельных токенов, а для интервалов, для которых впоследствии вычисляется вероятность соответствия сущностям. Путем перераспределения вероятностей сущностей из аннотированных интервалов к соседним авторы добились лучших результатов при решении задачи извлечения сущностей.

В системе HAMLET [11] для каждого кандидата предлагается вычислить ряд из 152 признаков и затем обучить классификатор бинарного дерева решений. Кандидаты определяются на основе их части речи, но закономерности вхождения слова в термин определяются автоматически на основе данных обучения.

Традиционно многие системы экстракции терминов основаны на гибридном подходе: сначала применяется лингвистическая фильтрация для определения синтаксически правдоподобных терминов-кандидатов, затем кандидаты оцениваются и классифицируются с использованием статистических функций, специальных метрик или машинного обучения.

Подход, примененный в проведенном эксперименте, опирается на методы машинного обучения, дополненные извлечением терминологии с помощью словаря и валидацией на основе правил.

Инструмент Terminator

Основная цель данной работы заключается в исследовании влияния языковых моделей на извлечение терминологии из текстов.

При разработке системы извлечения терминологии был применен метод на основе слабо контролируемого обучения. Этот метод, подробно описанный в статье [12], позволяет использовать словарь терминов для автоматической разметки текстов, а полученные тексты – для обучения модели, которая затем автоматически размечает другие тексты. Далее модель может быть обучена уже на объединении текстов. Таким образом, метод не требует больших затрат на подготовку данных для обучения.

Сама модель представляет собой дообученную языковую модель bert-base-multilingual-cased [13] для векторизации. В архитектуру также входят слой двунаправленной LSTM и два полносвязных слоя. Результаты работы нейронной сети валидируются эвристиками.

Недостатком указанного инструмента является то, что он плохо распознает точные границы термина из-за того, что модель была обучена на шумных данных.

Подготовка данных

Для экспериментов были использованы три набора данных:

– обучающее множество размеченных вручную текстов;

– обучающее множество автоматически размеченных текстов;

– тестовое множество текстов для получения метрик.

Обучающее множество размеченных вручную текстов состоит из 136 аннотаций научных работ по информационным технологиям на русском языке. В качестве сущностей были рассмотрены следующие типы терминов:

– термины, состоящие из одной или нескольких лексем («*граф*», «*датчик системы технического зрения*»);

– аббревиатуры («*МРТ*», «*СЛАУ*»);

– названия языков программирования («*Python*», «*C++*») и библиотек («*Pytorch*», «*SpaCy*»);

– понятия, разделенные дефисом, и термины с латинскими символами («*n-граммы*», «*BPE-токены*»).

Сущности были размечены в формате BIO (см. рисунок): каждому токену присваивается метка (тег) B-TERM, если сущность начинается с данного токена, I-TERM, если он находится внутри термина, или O, если он находится вне какого-либо объекта. Данный вид разметки не учитывает возможность перекрытия терминами друг друга, поэтому варианты вложенности терминов не были рассмотрены.

В результате проделанной работы подготовлен корпус текстов, содержащий разметку научных терминов, в ходе которой выделены 1 198 терминов и 12 809 токенов.

token	**tag**
В	O
статье	O
рассматриваются	O
перспективы	O
дальнейшего	O
роста	O
быстродействия	B-TERM
современных	O
корпоративных	B-TERM
информационных	I-TERM
систем	I-TERM

Пример таблицы, содержащей BIO-разметку

An example of a table containing BIO-markup

Обучающее множество автоматически размеченных текстов, состоящее из 1 764 текстов, было получено с использованием словаря терминов. Основная часть словаря была составлена разработчиками инструмента путем извлечения из Википедии всех названий статей, которые входят в подграф категории «Наука», и дальнейшего отбора [12]. В ходе работы полученный таким образом список из 17 252 терминов был дополнен еще 1 574 терминами.

В качестве тестовых данных, на которых фиксируются метрики моделей для дальнейшего сравнения, был использован корпус из 79 размеченных текстов. Тестовые данные не меняются для того, чтобы можно было провести адекватное сравнение результатов.

Все наборы данных находятся в открытом доступе (https://github.com/iis-research-team/ruserrc-dataset/tree/master/ruserrc-v2). Статистика для полученных данных приведена в таблице 1.

Таблица 1

Статистика для данных

Table 1

Data statistics

Множество	**Количество**		
	текстов	**токенов**	**терминов**
Обучающее, ручная разметка	136	12 809	2 028
Обучающее, автоматическая разметка	1 764	279 039	42 760
Тестовое	79	11 157	2 027

Исследуемые модели

Извлечение терминов из текстов научных статей можно рассматривать как задачу присвоения токенам аннотирующих меток. Предварительно обученные двунаправленные языковые модели демонстрируют высокие показатели качества решения подобных задач.

Известны два альтернативных подхода к предварительному обучению таких моделей – одноязычный и многоязычный. В то время как одноязычные модели показывают высокую производительность, многоязычные позволяют выполнять перевод с одного языка на другой и решать задачи для разных языков одновременно.

В данной работе сравниваются показатели качества извлечения терминологии из текстов на русском языке с помощью двух моделей – mBART и ruBERT.

Модель mBERT предварительно обучена на большом массиве необработанных многоязычных данных [13]. В процессе предварительного обучения модель решала две задачи: предсказание случайно замаскированного слова и предсказание следующего предложения. Таким образом, модель обучена выявлению закономерности строения предложений в языках из обучающего набора. Данную модель можно применять при обучении стандартного классификатора для решения таких задач, как классификация последовательностей или токенов.

Модель ruBERT (12 слоев, 768 скрытых слоев, 180 М параметров) была обучена на русскоязычной части Википедии и новостных данных [14]. На основе этих данных авторами был составлен словарь русских субтокенов. В качестве инициализации для данной модели использована мультиязычная версия BERT-базы.

Обе модели введены с помощью класса конфигурации BertForTokenClassification, имеющего последний линейный слой, что позволяет использовать модели для таких задач, как распознавание именованных сущностей (NER) или, что важно для данной работы, присвоение токенам аннотирующих меток.

Выбор языковой модели существенно сказывается на качестве работы всего инструмента, извлекающего терминологию, даже при применении совместно с другими методами. Важно понять, насколько сильно предварительное обучение на текстах конкретного языка влияет на полноту и точность результатов выявления терминов в текстах на этом языке.

Описание эксперимента

Для сравнения качества извлечения терминов из текстов научных статей двумя вышеуказанными языковыми моделями было проведено несколько этапов исследования.

Каждая модель была дообучена на размеченном вручную обучающем наборе текстов. После этого зафиксированы и представлены для сравнения числовые показатели качества результатов извлечения терминологии только моделью, работы модели с применением валидации эвристиками, работы модели, подкрепленной словарем и эвристиками.

Те же метрики были зафиксированы для данных моделей, обученных на автоматически размеченном корпусе текстов.

Первый этап: применение моделей без подключения иных средств. Методика эксперимента следующая. Сначала для дообучения модели были использованы 136 вручную размеченных текстов. Оптимальная скорость обучения (learning rate) выбрана равной 10-6, а размер пакета (batch size) – 12. В процессе подбора параметров было выяснено, что именно такие значения позволяют предотвратить переобучение модели и получить при этом лучший результат. На данном этапе метрики показывают, что при частичном совпадении обе модели дают одинаковую производительность, а при точном совпадении модель, предварительно обученная на русскоязычных текстах, дает лучшие результаты.

Затем были добавлены 1 764 автоматически размеченных текста. Так как проверка качества разметки добавленной части корпуса не проводилась, даже при его сравнительно большом размере метрики, полученные для моделей, которые обучались на нем, оказались ниже, чем для тех же моделей, обученных на аннотированных вручную текстах.

Второй этап: ансамбль моделей и эвристик. Безусловно, результаты предсказаний модели нельзя считать точным списком извлеченных терминов. Выделенные моделью цепочки токенов могут включать в себя слова, не относящиеся к термину, служебные слова, случайные символы или вовсе не являться терминами. Поэтому для улучшения качества извлечения терминов требуется валидация извлеченных кандидатов эвристиками: исключение заведомо ложных случаев выделения терминов и уточнение границ.

Задача определения границ терминов более неоднозначна и вызывает больше ошибок, чем отнесение слова к типу «термин». Границы упоминания термина могут быть неоднозначными из-за пограничных слов (например, предлогов или союзов). В целях совершенствования инструмента были написаны и применены эвристики для обработки таких случаев, как валидация последовательности токенов между символами открывающих и закрывающих кавычек, а также исправление разметки терминов, которые пишутся через дефис, и некоторые другие.

Результаты эксперимента показали, что эвристики действительно улучшили точность извлечения терминов при полном совпадении, то есть уменьшилось число ложных предсказаний моделей. Стоит отметить, что на частичном совпадении некоторые результаты оказались ниже, чем результаты работы тех же моделей без применения к ним эвристик. Это объясняется тем, что большая часть эвристик исправ-

ляет ошибки первого рода, то есть убирает метки B-TERM или I-TERM у тех токенов, которые не должны их иметь, но не выполняет присвоение этих меток.

Третий этап: ансамбль модели, словаря и эвристик. На третьем этапе система извлекает термины с помощью не только обученной модели, но и описанного выше словаря. Также к результатам применяются эвристики.

Первым шагом эксперимента на данном этапе можно считать дообучение моделей с помощью размеченного вручную обучающего множества (136 текстов) и получение метрик в трех вариантах.

Далее, после получения 1 764 автоматически аннотированных текстов, модели были дообучены на втором обучающем множестве. Проведено сравнение результатов работы моделей изолированно и в ансамбле со словарем и эвристиками.

Полученные результаты сравнивались с эталоном – тестовым корпусом (79 текстов). Все полученные метрики отражены в таблице 2.

Результаты экспериментов

В конце каждого этапа эксперимента фиксировались показатели качества работы инструмента на тестовом множестве текстов (79 текстов). Для оценки качества алгоритма использовались следующие метрики:

– точность, полнота и F1-мера частичного совпадения – проверка того, что модель в принципе смогла определить, относится данный токен к термину или нет (без уточнения его границ);

– точность, полнота и F1-мера полного совпадения – проверка правильности присвоения токенам моделью всех трех типов меток.

Для сравнения в первой строчке таблицы 2 приведены результаты работы инструмента до проведенного эксперимента. Эти результаты получены с использованием мультиязычной модели mBERT, дообученной на текстах, размеченных автоматически, в объединении со словарем и эвристиками.

В таблице также представлены численные оценки качества извлечения терминологии каждой моделью в трех вариантах: только дообученная модель, дообученная модель, к которой применены эвристики, и модель, подкрепленная эвристиками и словарем.

Полужирным шрифтом выделены лучшие показатели моделей на каждом обучающем множестве. Обе модели качественно решают задачу распознавания терминов, что видно по хорошим результатам на частичном совпадении. Мультиязычная модель mBERT доста-

Таблица 2

Полученные метрики

Table 2

Received metrics

Данные	Эксперимент	Полное совпадение			Частичное совпадение		
		Точность	Полнота	F1	Точность	Полнота	F1
Автоматическая разметка	mBERT + словарь + эвристики	0.40	0.30	0.34	0.77	0.75	0.76
Обучающее множество с ручной разметкой	mBERT	0.40	0.46	0.43	**0.89**	**0.88**	**0.88**
	mBERT + эвристики	0.49	0.45	0.47	0.86	0.86	0.86
	mBERT + словарь + эвристики	0.47	0.50	0.48	0.86	0.87	0.87
	ruBERT	0.48	0.50	0.49	**0.89**	**0.88**	**0.88**
	ruBERT + эвристики	**0.52**	0.47	0.49	**0.89**	**0.88**	**0.88**
	ruBERT + словарь + эвристки	0.49	**0.51**	**0.50**	0.86	0.87	0.87
Обучающее множество с автоматической разметкой	mBERT	0.29	0.31	0.30	**0.80**	0.74	0.74
	mBERT + эвристики	**0.42**	0.38	**0.40**	**0.80**	0.79	0.79
	mBERT + словарь + эвристики	0.41	**0.39**	**0.40**	**0.80**	**0.80**	**0.80**
	ruBERT	0.32	0.34	0.33	0.78	0.76	0.75
	ruBERT + эвристики	0.40	0.37	0.38	0.79	0.79	0.79
	ruBERT + словарь + эвристкии	0.38	**0.39**	0.38	0.79	0.74	0.76

точно хорошо выявляет термины, но неточно определяет их границы. Это следует из ее изначальной многофункциональности. Наилучший результат при полном совпадении принадлежит модели ruBERT, подкрепленной словарем и эвристиками. Результаты моделей на полном совпадении ожидаемо ниже, чем на частичном совпадении, что вновь обращает внимание на задачу определения границ терминов в текстах.

Можно заметить, что модель ruBERT, дообученная на аннотированном вручную корпусе текстов, извлекает термины в половине случаев, что является лучшим результатом полного совпадения. На наборе автоматически размеченных данных предложенный комбинированный метод, включающий использование словаря, модели и эвристик, дает хорошие результаты для многоязычного mBERT как для полного, так и для частичного совпадений. Однако разница показателей предсказаний моделей на размеченном вручную корпусе и на корпусе с автоматической разметкой позволяет сделать вывод о том, что качество разметки обучающего множества влияет на качество дальнейшей работы инструмента. Интересно, что для многоязычной модели подключение словаря увеличивает полноту предсказаний модели и на полном, и на частичном совпадении, а для моноязычной, ruBERT, на частичном совпадении полнота уменьшается. То есть модель стала присваивать метки -TERM точнее, но реже.

Сравнивая результаты работы системы на двух корпусах, можно сделать вывод, что качество разметки обучающего набора существенно влияет на качество извлечения терминологии. При дообучении на размеченных вручную текстах модели дали лучшие результаты, чем при дообучении на автоматически размеченном наборе, даже несмотря на то, что объемы последнего в несколько раз превосходят объемы первого обучающего набора.

Заключение

В работе подробно рассмотрены основные теории и подходы к извлечению предметной терминологии из текстов научных статей. Проведены эксперименты и дальнейшее сравнение результатов интеграции различных языковых моделей в инструмент, извлекающий терминологию. Для сравнения выбраны модель mBERT, предобученная на текстах на большом количестве языков, и ruBERT, предобученная только на русскоязычных данных. Модернизирован алгоритм извлечения терминологии из текстов, изучено влияние выбора языковой модели на полноту и точность результата.

Извлеченные термины могут быть использованы в качестве элементов тематического указателя или ключевых слов для автоматического индексирования документов. Также автоматическое извлечение терминов из научных текстов упрощает процесс создания словаря терминов или составления онтологии для определенной предметной области, что особенно актуально для быстро меняющихся отраслей.

Литература

1. Dowlagar S., Mamidi R. Unsupervised technical domain terms extraction using term extractor. Proc. XVII ICON, 2020, pp. 5–8.

2. Stoykova V., Stankovic R. Using query expansion for cross-lingual mathematical terminology extraction. In: AISC, 2019, pp. 154–164. DOI: 10.1007/978-3-319-91189-2_16.

3. Das B., Pal S., Mondal S., Dalui D., Shome S.K. Automatic keyword extraction from any text document using N-gram rigid collocation. IJSCE, 2013, vol. 3, no. 2, pp. 238–242.

4. Sha W., Hua B., Linqi S. A Pattern and POS auto-learning method for terminology extraction from scientific text. Data and Information Management, 2021, vol. 5, no. 3, pp. 329–335. DOI: 10.2478/dim-2021-0005.

5. Quirchmayr T., Paech B., Kohl R. et al. Semi-automatic rule-based domain terminology and software feature-relevant information extraction from natural language user manuals. Empirical Software Engineering, 2018, vol. 23, no. 6, pp. 3630–3683. DOI: 10.1007/s10664-018-9597-6.

6. Stanković R., Krstev C., Obradovic I., Lazic B., Trtovac A. Rule-based automatic multi-word term extraction and lemmatization. Proc. X Int. Conf. LREC, 2016, pp. 507–514.

7. Saraiva F.Z.R., da Silva T.L.C., de Macêdo J.A.F. Aspect term extraction using deep learning model with minimal feature engineering. In: AISE, 2020, vol. 12127, pp. 185–198. DOI: 10.1007/978-3-030-49435-3_12.

8. Lopez P., Romary L. HUMB: Automatic key term extraction from scientific articles in GROBID. Proc. V Int. Workshop on Semantic Evaluation, 2010, pp. 248–251.

9. Conrado M., Pardo T., Rezende S. O. A machine learning approach to automatic term extraction using a rich feature set. Proc. NAACL, 2013, pp. 16–23.

10. Zhou K., Li Y., Li Q. Distantly supervised named entity recognition via confidence-based multi-class positive and unlabeled learning. Proc. LX Annual Meeting of the Association for Computational Linguistics, 2022, vol. 1, pp. 7198–7211.

11. Terryn A.R., Hoste V., Lefever E. HAMLET. Terminology, 2021, vol. 27, no. 2, pp. 254–293. DOI: 10.1075/term.20017.rig.

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

13. Devlin J., Chang M.W., Lee K., Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. Proc. NAACL, 2018, vol. 15, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

14. Kuratov Y., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language. Proc. Int. Conf. Dialogue, 2019, pp. 333–339.

Software & Systems
DOI: 10.15827/0236-235X.140.689-697

Received 02.07.22, Revised 24.08.22
2022, vol. 35, no. 4, pp. 689–697

Terms extraction from texts of scientific papers

Ya.Yu. Dementeva [1], *Student, y.dementeva@g.nsu.ru*
E.P. Bruches [1, 2], *Junior Researcher, Senior Lecturer, bruches@bk.ru*
T.V. Batura [2], *Ph.D. (Physics and Mathematics), Associate Professor, Senior Researcher, tatiana.v.batura@gmail.com*

[1] *Novosibirsk State University, Novosibirsk, 630090, Russian Federation*
[2] *A.P. Ershov Institute of Informatics Systems SB RAS, Novosibirsk, 630090, Russian Federation*

Abstract. The relevance of the task of extracting terms from the texts of scientific articles is due to the need for automatic annotation and extracting keywords in an ever-increasing flow of scientific and technical documents. This paper explores the influence of various language models on the quality of extracting scientific terms from Russian texts.

We compare two models: the mBERT model that was pretrained on texts of different languages, and the ruBERT model pretrained only on Russian data. Two training sets of annotated texts were prepared. The authors carried out fine-tuning and further comparison of the performance indicators of the two models using these training sets. They also studied the influence of the choice of the language model on the quality of extracting the terminology contained in the texts of scientific articles.

The results have become the base for modernizing the algorithm for extracting terminology from texts applied by the Terminator tool, developed at the A.P. Ershov Institute of Informatics Systems.

The obtained results showed that within the framework of the task of extracting terminology from the texts of Russian scientific articles, the ruBERT model, which gave the best performance in an ensemble with a dictionary and heuristics, can be considered as the most applicable model. In addition, the difference in the results of models on full and partial match can be stated due to the problem of defining the boundaries of terms in the texts described in the paper.

The results obtained also allow concluding that the quality of the training set markup affects the quality of terminology extraction.

Keywords: terminology extraction, natural language processing, machine learning, language model, mBERT, ruBERT, terms dictionary.

References

1. Dowlagar S., Mamidi R. Unsupervised technical domain terms extraction using term extractor. *Proc. XVII ICON*, 2020, pp. 5–8.

2. Bruches E.P., Batura T.V. Method for automatic term extraction from scientific articles based on weak supervision. *Vestn. NSU. Ser.: Information Technologies*, 2021, vol. 19, no. 2, pp. 5–16. DOI: 10.25205/1818-7900-2021-19-2-5-16 (in Russ.).

3. Stoykova V., Stankovic R. Using query expansion for cross-lingual mathematical terminology extraction. In: *AISC*, 2019, pp. 154–164. DOI: 10.1007/978-3-319-91189-2_16.

4. Das B., Pal S., Mondal S., Dalui D., Shome S.K. Automatic keyword extraction from any text document using N-gram rigid collocation. *IJSCE*, 2013, vol. 3, no. 2, pp. 238–242.

5. Sha W., Hua B., Linqi S. A Pattern and POS auto-learning method for terminology extraction from scientific text. *Data and Information Management*, 2021, vol. 5, no. 3, pp. 329–335. DOI: 10.2478/dim-2021-0005.

6. Quirchmayr T., Paech B., Kohl R. et al. Semi-automatic rule-based domain terminology and software feature-relevant information extraction from natural language user manuals. *Empirical Software Engineering*, 2018, vol. 23, no. 6, pp. 3630–3683. DOI: 10.1007/s10664-018-9597-6.

7. Stanković R., Krstev C., Obradovic I., Lazic B., Trtovac A. Rule-based automatic multi-word term extraction and lemmatization. *Proc. X Int. Conf. LREC*, 2016, pp. 507–514.

8. Saraiva F.Z.R., da Silva T.L.C., de Macêdo J.A.F. Aspect term extraction using deep learning model with minimal feature engineering. In: *AISE*, 2020, vol. 12127, pp. 185–198. DOI: 10.1007/978-3-030-49435-3_12.

9. Lopez P., Romary L. HUMB: Automatic key term extraction from scientific articles in GROBID. *Proc. V Int. Workshop on Semantic Evaluation*, 2010, pp. 248–251.

10. Conrado M., Pardo T., Rezende S. O. A machine learning approach to automatic term extraction using a rich feature set. *Proc. NAACL*, 2013, pp. 16–23.

11. Zhou K., Li Y., Li Q. Distantly supervised named entity recognition via confidence-based multi-class positive and unlabeled learning. *Proc. LX Annual Meeting of the Association for Computational Linguistics*, 2022, vol. 1, pp. 7198–7211.

12. Terryn A.R., Hoste V., Lefever E. HAMLET. *Terminology*, 2021, vol. 27, no. 2, pp. 254–293. DOI: 10.1075/term.20017.rig.

13. Devlin J., Chang M. W., Lee K., Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proc. NAACL*, 2018, vol. 15, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

14. Kuratov Y., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language. *Proc. Int. Conf. Dialogue*, 2019, pp. 333–339.

Для цитирования

Дементьева Я.Ю., Бручес Е.П., Батура Т.В. Извлечение терминов из текстов научных статей // Программные продукты и системы. 2022. Т. 35. № 4. С. 689–697. DOI: 10.15827/0236-235X.140.689-697.

For citation

Dementeva Ya.Yu., Bruches E.P., Batura T.V. Terms extraction from texts of scientific papers. *Software & Systems*, 2022, vol. 35, no. 4, pp. 689–697 (in Russ.). DOI: 10.15827/0236-235X.140.689-697.