УДК 519.688; 004.89 doi: 10.15827/0236-235X.149.005-016 2025. Т. 38. № 1. С. 5–16

Методы и средства извлечения терминов из текстов для терминологических задач

Е.И. Большакова [1]✉, В.В. Семак [1]

[1] Московский государственный университет им. М.В. Ломоносова, г. Москва, 119991, Россия

Ссылка для цитирования
Большакова Е.И., Семак В.В. Методы и средства извлечения терминов из текстов для терминологических задач // Программные продукты и системы. 2025. Т. 38. № 1. С. 5–16. doi: 10.15827/0236-235X.149.005-016
Информация о статье
Группа специальностей ВАК: 1.2.3
Поступила в редакцию: 18.08.2024 После доработки: 14.09.2024 Принята к публикации: 29.09.2024

Аннотация. В статье рассматривается современное состояние области автоматического извлечения терминов из специализированных текстов на естественном языке, включая научно-технические документы. К числу актуальных практических приложений методов и средств извлечения терминов из текстов относятся создание терминологических словарей, тезаурусов и глоссариев предметных областей, а также выявление ключевых слов и построение предметных указателей для узкоспециализированных документов. Представлен обзор подходов к автоматическому распознаванию и извлечению терминологических слов и словосочетаний, охватывающих традиционные статистические методы, а также методы на основе машинного обучения, включая обучение по признакам терминов и с применением современных нейросетевых языковых моделей. Проводится сравнение подходов, в том числе оценки качества распознавания и извлечения терминов, указываются наиболее известные программные средства для автоматизации извлечения терминов в рамках статистического подхода и обучения по признакам. Описываются исследования, проведенные авторами для распознавания терминов на базе нейросетевых языковых моделей применительно к обработке научных текстов по математике и программированию на русском языке. Кратко характеризуется набор данных с терминологической разметкой, созданный для обучения программных моделей распознавания терминов и охватывающий данные семи близких предметных областей. Программные модели разработаны на основе предобученной нейросетевой модели BERT с ее дообучением двумя способами: как бинарного классификатора терминов-кандидатов (предварительно извлеченных из текстов) и как классификатора для последовательной разметки терминологических слов в обрабатываемых текстах. Для разработанных моделей экспериментально определены оценки качества распознавания терминов, проведено сравнение со статистическим методом. Лучшие результаты продемонстрировали модели бинарной классификации, существенно превосходя остальные рассмотренные подходы. Проведенные эксперименты показывают применимость обученных моделей и для текстов близкой научной области.

Ключевые слова: автоматическая обработка текстов специализированных областей, терминологические слова и словосочетания, автоматическое извлечение терминов, машинное обучение для распознавания терминов, извлечение терминов на базе языковой модели BERT

Введение. Автоматическое извлечение терминов из текстов (*Automatical Term Extraction*, ATE, или *Automatical Term Recognition*, ATR) – одна из традиционных задач в области автоматической обработки текстов [1, 2]. Ее актуальность обусловлена стремительным развитием науки и техники и появлением в связи с этим новых терминов – слов и словосочетаний, обозначающих понятия разных предметных областей. Хотя задача ATE изучается более трех десятилетий, ее решения не достигают высокого качества, как во многих других задачах автоматической обработки текстов. Прежде всего это объясняется сложностью и неоднозначностью критерия терминологичности, который относится к области семантики и трудно формализуется и применяется на практике человеком.

В то же время терминология специальных предметных областей представляет собой важный пласт знаний, необходимых при решении многих прикладных задач. Методы ATE применяются для автоматизированного построения

– терминологических словарей, тезаурусов и онтологий предметных областей по коллекциям специализированных текстов [3];

– глоссариев – перечней терминов с их определениями для проблемно-ориентированных областей и узкоспециализированных документов [4, 5];

– предметных указателей для научно-технических книг и документов [6].

Задача ATE также важна для улучшения методов машинного перевода специализированных текстов [7].

К настоящему моменту известны следующие подходы к автоматическому распознаванию и извлечению терминов:

– ставший традиционным статистический подход, основанный на ранжировании терминов-кандидатов по терминологичности с помощью статистических мер и применяемый в основном для обработки текстовых коллекций [8–10];

– подход на основе машинного обучения бинарного классификатора термин/нетермин (для предварительно выделенных кандидатов) [11, 12], позволяющий выявить наилучшую комбинацию признаков (лингвистических и статистических) для распознавания терминов;

– подход последовательной разметки слов текста, при котором на базе машинного обучения строится модель для распознавания терминов прямо в тексте, путем выявления и разметки входящих в термины слов [13];

– подход с применением современных нейросетевых трансформерных языковых моделей (например, BERT [14]) для машинного обучения, при котором вместо набора признаков для обучения выступает контекстная информация, воплощенная в векторах слов (эмбеддингах) применяемой языковой модели [15–18].

Последний подход применяется либо для бинарной классификации термин/нетермин, либо для последовательной разметки терминов в обрабатываемых текстах. В ряде работ он продемонстрировал лучшие результаты решения задачи ATE (предположительно, за счет глубокого предобучения применяемых нейросетевых языковых моделей на больших массивах неразмеченных текстов), однако требует дальнейшего исследования, поскольку эксперименты были относительно немногочисленны и проводились лишь для некоторых языков текста (преимущественно английского). Одна из возникающих проблем связана с недостатком отрытых текстовых корпусов с эталонной терминологической разметкой, необходимых для проведения экспериментов по обучению моделей ATE и их оценки. Особенно острая ситуация сложилась для русского языка, работ по применению указанного подхода практически нет.

Отдельной темой для исследований является применимость моделей ATE, обученных для текстов определенной предметной области, для извлечения терминов в текстах из другой области без потери качества, которая обычно бывает при машинном обучении на признаках.

В настоящей работе задача ATE рассматривается применительно к русскому языку в рамках подхода на базе нейросетевых языковых моделей. Основная цель – экспериментально оценить и сравнить два способа распознавания терминов на основе нейросетевой модели BERT:

– бинарный классификатор для распознавания терминов с использованием контекста в виде объемлющего предложения;

– машинный классификатор, распознающий термины в тексте путем последовательной разметки входящих в них слов.

Обучение классификаторов, их оценка и сравнение проводились на одном и том же размеченном наборе данных, созданном для данной задачи из русскоязычных учебно-научных текстов. Набор данных включал термины из нескольких областей математики и программирования, что позволило оценить качество моделей при переходе от одной области к другой (от математики к программированию и наоборот). Дополнительно сопоставлялось качество извлечения терминов указанными классификаторами (стандартные метрики точности, полноты, F-меры) и методом статистического подхода. Показано, что подход на базе модели BERT как бинарного классификатора термин/нетермин достигает 73 % F1-меры и по качеству распознавания терминов превосходит модель последовательной разметки, а также статистический метод.

Статистический подход к извлечению терминов

Традиционные статистические методы ATE [8–10] хорошо изучены и основаны на предположении, что термины часто встречаются в текстах в определенных грамматических формах, для распознавания которых используются статистические и лингвистические свойства (признаки) терминов.

К лингвистическим признакам в первую очередь относятся грамматические образцы многословных терминов, например, прилагательное + существительное + существительное (*спектральный коэффициент излучения*) и др. Лингвистические свойства могут учитывать употребление терминов в определенных контекстах (например, «*...будем называть сюръекцией такое отображение...*»), а также списки стоп-слов − слов и словосочетаний, которые не могут быть терминами или входить в них. К ним относятся некоторые слова общей лексики (*другой, схема* и т.п.) и оценочные слова (*плохой* и т.д.).

Статистические признаки (меры) основаны на частотах употребления слов в текстах и оценивают вероятность того, что те или иные слова и словосочетания на самом деле будут терминами. Одним из наиболее эффективных критериев этой группы является *C-value* [19] – мера терминологичности для многословных терминов, учитывающая частоту словосочетания, число входящих в него слов, частоту вложенных в него словосочетаний и поощряющая словосочетания, не входящие в состав других, более длинных.

Общая схема всех статистических методов извлечения терминов из текста включает следующие этапы:

– распознавание слов и словосочетаний по заданным грамматическим образцам и контекстам; извлечение из текста распознанных единиц в качестве терминов-кандидатов;

– ранжирование этих кандидатов по значению выбранной статистической меры с целью получения истинных терминов в верхней части ранжированного списка;

– фильтрация (удаление) некоторых терминов-кандидатов с учетом заданного списка стоп-слов;

– извлечение первых K элементов из ранжированного и отфильтрованного списка, которые и считаются терминами.

Оценка качества извлечения терминов в рамках рассматриваемого подхода обычно осуществляется с использованием метрик точности и средней точности. Точность равна доле истинных терминов среди всех извлеченных элементов (*Precision@K* – точность на *K* кандидатах), а средняя точность (*Average Precision, AP*) оценивает, насколько истинные термины находятся ближе к началу ранжированного списка:

$$Precision@K = \frac{\sum_{i=1}^{k} rel(i)}{k},$$

$$AP = \frac{\sum_{k=1}^{n} Precision@K \times rel(k)}{R},$$

где $rel(i) = 1$, если термин-кандидат под номером i действительно является термином, и $rel(i) = 0$ в противном случае; R – число истинных терминов среди извлеченных.

Эффективность статистических методов ATE зависит от предметной области обрабатываемых текстов, размера текстовой коллекции и параметров метода (в частности, от набора грамматических образцов), на практике это обычно 30–60 % средней точности. В силу своей статистической природы такие методы лучше работают для объемных текстов и обычно применяются для извлечения терминологии из коллекций проблемно-ориентированных текстов, хотя могут быть применимы и для отдельных текстов [20], например, для построения глоссариев и предметных указателей [21]. Для повышения качества извлечения терминов в рамках статистического подхода применяются сложные эвристические и зависящие от области текстов стратегии фильтрации, например, такая стратегия представлена в работе [6], где средняя точность извлечения терминов достигала 70 %. Преимущество подхода в том, что статистические методы не требуют каких-либо размеченных данных.

В рамках рассмотренного подхода созданы программные инструменты, например [22, 23], отличающиеся языком программирования (в частности, Java в [24], Scala в [25, 26]) и позволяющие задавать параметры реализуемого статистического метода – статистические меры, набор грамматических образцов, список стоп-слов, а также вычислять значения выбранных мер на заданных текстовых коллекциях.

Машинное обучение на основе набора признаков терминов

При этом подходе для задачи ATE применяется традиционное машинное обучение (байесовский метод, логистическая регрессия, метод опорных векторов, градиентный бустинг и др.) и за счет обучения на размеченных данных определяется значимость признаков, по которым выносится решение, является ли слово или словосочетание термином [3, 11, 12].

Общая схема извлечения терминов из текста включает три этапа.

1. Формирование набора терминов-кандидатов. Как и в статистическом подходе, оно обычно выполняется с помощью грамматических образцов, но в качестве кандидатов могут выступать и всевозможные N-граммы слов текста.

2. Вычисление значений признаков этих кандидатов: орфографических (регистр букв, наличие небуквенных символов и др.), статистических (статистические меры), лингвистических и контекстных (наличие определенных слов в самом кандидате или в его контексте и др.).

3. Обучение (а затем и применение) машинного классификатора для распознавания терминов на основе вычисленных признаков.

К примеру, на этапах 1 и 2 в работе [11] учитывались *N*-граммы (*N* от 1 до 5), за исключением стоп-слов, и такие их признаки, как частотность и C-value, а в работе [3] применялись грамматические образцы и несколько статистических признаков.

Машинное обучение позволяет выявить оптимальную комбинацию признаков терминов и тем самым повысить качество их извлечения. Качество традиционно оценивается как точность (доля истинных терминов среди найденных), полнота (доля истинных терминов среди всех терминов текста) и F1-мера (среднее значение гармонической полноты и точности).

Для обучения классификатора необходим корпус с терминологической разметкой, что долгое время составляло проблему из-за малочисленности и отсутствия общепринятых корпусов с ручной (эталонной) разметкой. В исследовательских работах последних лет из немногих доступных ныне корпусов с ручной разметкой терминов преимущественно используется ACTER [27] – недавно созданный мультиязычный корпус, включающий тексты на трех языках (английском, французском, голландском) для четырех предметных областей (ветровая энергетика, сердечная недостаточность, коррупция, выездка лошадей). Кроме размеченных текстов, корпус содержит списки терминов для каждого языка и каждой области.

В работе [12] на основе корпуса ACTER проведено сравнение статистического метода TermoStat (грамматические образцы и статистические меры) с моделью машинного обучения HAMLET. При обучении HAMLET был применен метод случайного леса и рассмотрены более 130 различных признаков терминов. Для различных комбинаций языков и предметных областей обучены отдельные классификаторы, усредненное значение F1-меры для них оказалось в пределах 46.7–54.9 % , что значительно выше, чем полученные 28 % F1-меры, продемонстрированной моделью TermoStat.

Хотя в ряде работ показано, что машинное обучение на признаках достигает лучших результатов при извлечении терминов для заданной предметной области, ценность подхода ограничена, поскольку качество работы обученных классификаторов обычно падает (и даже существенно) на текстах из другой области, где термины могут иметь другие значимые признаки. Одна из причин связана с тем, что набор признаков слабо отражает семантику текста, аналогичная причина действует и для статистического подхода. Еще одним слабым местом машинного обучения на признаках является необходимость ручного (экспертного) подбора признаков.

Распознавание терминов на основе языковых моделей

С появлением в практике обработки текстов нейросетевых языковых моделей архитектуры Transformer, таких как BERT [14], в ряде работ по ATE [15–18] было предложено использовать для машинного обучения вместо наборов признаков терминов (лингвистических, статистических, контекстных) векторные представления слов из предобученных языковых моделей − контекстуализированные эмбеддинги, сохраняющие контекст применения слов.

В работах [15, 16] предобученные модели BERT дообучались как бинарные классификаторы предсказывать для заданной пары из предложения и *N*-граммы из него (рассматриваемой как кандидат в термины), является ли *N*-грамма термином или нет. Дообучение (*fine-tuning*) проводилось на размеченных данных по аналогии с задачей предсказания следующего предложения: если *N*-грамма являлась термином в контексте предложения, этот пример пары являлся положительным, в противном случае − отрицательным. При таком подходе обучающие данные должны содержать набор пар вида <предложение текста + термин-кандидат из него> и термины-кандидаты заранее извлекаются из текста (как и в вышеописанных подходах к ATE).

В статье [15] проведено сравнение дообученных как бинарные классификаторы моделей BERT (RoBERTa для английского языка и CamemBERT для французского) с классификатором на основе метода XGBoost, обученным на наборе признаков терминов (лингвистических и статистических). Положительные примеры обучающего набора были построены с использованием данных корпуса ACTER, тогда как отрицательные пары (*N*-граммы, которые не являются терминами) были сгенерированы случайным образом. В экспериментах обученная модель XGBoost показала высокую точность распознавания терминов, но низкую полноту, в итоге дав около 27 % F1-меры, в то время как модель классификации на основе BERT существенно превзошла этот результат, показав 48 % F1-меры.

Отметим, что, хотя описанный подход к бинарной классификации термин/нетермин на базе моделей BERT не требует ручного подбора признаков, у него есть существенный недостаток: для получения обучающего набора данных необходима генерация пар с участием всех возможных *N*-грамм, что вычислительно затратно.

Машинное обучение для распознавания терминов на основе последовательной разметки

Другой способ распознавания терминов на базе нейросетевых языковых моделей, представленный в работах [13, 17], не требует предварительного извлечения кандидатов в термины, вместо этого термины распознаются прямо в текстах моделью машинного обучения. Модель обучается на тексте с размеченными терминами и затем выполняет последовательную разметку слов-токенов текста, действуя аналогично таким известным задачам разметки последовательностей, как определение части речи слов и распознавание именованных сущностей. Точнее, обучается модель классификации, которая для каждого слова-токена текста предсказывает, является ли токен частью какого-либо термина или нет. Для этого используется разметка токенов BIO или IO (B помечает начальное слово термина, I − его внутренние слова, O − слова текста, не являющиеся частью никакого термина). Обученный классификатор проставляет эти метки словам текста, после чего предсказанные метки могут быть использованы для извлечения терминов из размеченного текста и формирования из них списка.

В работе [13] для обучения классификаторов применялись данные из мультиязычного корпуса ACTER, сравнивались несколько методов машинного обучения: часто применяемый для последовательной разметки метод CRF (*Conditional Random Field*), выполняющий обучение на признаках, рекуррентные нейронные сети (RNN) с векторными представлениями слов (эмбеддингами) и дообучение модели BERT как классификатора токенов. При этом рассматривались мультиязычные и одноязычные эмбеддинги моделей BERT для представленных в корпусе языков. Эксперименты показали, что модель RNN с одноязычными эмбеддингами достигает 47−57 % F1-меры для распознавания терминов (в зависимости от предметной области текстов), превосходя таковые оценки CRF-модели и дообученных моделей BERT. Мультиязычные эмбеддинги могут даже улучшить F1-меру до 75 %, если дополнительно к обучающим данным для целевой предметной области берутся данные на другом языке, но для той же области (однако на практике наборы данных с терминологической разметкой на нескольких языках встречаются крайне редко).

Исследование подхода к ATE на основе последовательной разметки, но для словенского языка описано в работе [17], в ней применялись несколько моделей семейства BERT и недавно созданный размеченный корпус RSDO5 с терминами четырех предметных областей (биомеханика, химия, ветеринария, лингвистика). Были реализованы 12 моделей распознавания терминов с обучением на текстах одной области и тестированием на другой, результаты показали высокое значение F1-меры − 64−71 % , что доказывает возможность переноса обученных моделей с одной области на другую.

В статье [16] проведено сравнение двух подходов на базе нейросетевых языковых моделей: последовательной разметки токенов текста и бинарной классификации для предсказания термин/нетермин по парам <предложение + термин-кандидат>. На данных корпуса ACTER были проведены эксперименты с кроссязыковым обучением мультиязычной модели XML-RoBERTa (семейства BERT), то есть с обучением на одном языке и тестированием на другом, при этом рассматривались разные варианты смены области текста для обучения и тестирования. Для разных пар языков и пар областей обученный бинарный классификатор показал невысокие результаты: 40−58 % F1-меры, в то время как классификатор для последовательной разметки продемонстрировал 44−69 %.

Похожее исследование представлено в [18], где также описаны эксперименты в условиях смены предметной области текста для обучения и тестирования для текстов и предметных областей корпуса ACTER. Однако обученные классификаторы BERT как для последовательной разметки, так и для бинарной классификации термин/нетермин показали довольно низкие результаты: в пределах 34−43 % F1-меры в зависимости от конкретной пары областей и рассматриваемого языка текста.

Таким образом, в рассмотренных работах в области ATE на основе нейросетевых языковых моделей оценки качества распознавания терминов обученными моделями отличаются, варьируясь в зависимости от языка текстов и применяемых для обучения данных, что требует дальнейшего изучения. Тем не менее качество распознавания превосходит таковое для статистических методов, а ряд обученных моделей показал довольно высокое значение F1-меры распознавания терминов. Для русского языка подобные модели бинарной классификации и последовательной разметки текстов практически не исследованы из-за отсутствия подходящих размеченных наборов данных, и настоящая работа восполняет этот пробел, включая вопрос о переносе моделей, обученных для одной предметной области, на другую область.

Данные для обучения нейросетевых моделей извлечения терминов для русского языка

Для разработки и экспериментальной оценки методов распознавания терминов на основе нейросетевых языковых моделей применительно к текстам на русском языке был построен набор размеченных данных.

В качестве исходных специализированных текстов были выбраны научные тексты по математике и программированию, поскольку задача ATE особенно актуальна для обработки научно-технических текстов, содержащих много специальных терминов. Коллекция текстов включала свободно доступные тексты семи учебных пособий на русском языке общим объемом 267 тысяч слов-токенов. Важно, что для этих текстов по результатам предшествующих исследований [21] были известны списки входящих в них (извлеченных) терминов. Тексты охватывали такие предметные области, как *математический анализ* (МатАн), *дифференциальные уравнения* (ДифУр), *дискретная математика* (ДисМат), *искусственный интеллект* (ИИ), *формальные грамматики* (ФормГр), *системы программирования* (СисПрог) и *языки программирования* (ЯзПрог). Эти области, хотя имеют некоторое число общих терминов (например, *функция, множество* и др.), все же существенно различаются по терминологии. Статистика по объему текстов (число текенов-слов) и числу уникальных (разных) терминов в них представлена в таблице 1 (вторая и третья строки).

Для обучения бинарного классификатора пар вида <предложение + словосочетание из него>, который определяет, является ли термином данное словосочетание, был разработан (на языке Python, с помощью библиотеки SpaCy) соответствующий набор с более чем 23 тысячами положительных и отрицательных примеров пар. Предложения брались из текстов коллекции, и положительные примеры включали термин из них, а отрицательные − словосочетания из этих предложений, которые не были терминами. В третьей строке таблицы 1 представлено количество примеров-пар по областям. Приведем положительный пример: <*для бесконечных множеств говорить о количестве элементов не имеет смысла, но говорить о мощности множества можно* + *мощности множества*> и отрицательный: <*задача о нахождении кратчайшего расстояния может быть решена прямым перебором всевозможных расстояний* + *всевозможных расстояний*>.

При построении набора примеров все тексты были сегментированы на предложения. Для формирования положительных примеров найдены все вхождения каждого термина в предложения текстов коллекции, и предложения связаны с терминами, входящими в них. Для составления отрицательных примеров в текстах найдены все N-граммы длиной менее пяти и состоящие только из существительных и прилагательных, из которых отброшены все термины, а оставшиеся N-граммы (например, *данное решение*) связаны с исходными предложениями. Количество отрицательных примеров было уравнено с числом положительных для баланса положительного и отрицательного классов.

Таблица 1

Статистика по областям текстовой коллекции

Table 1

Statistics on text collection domains

Область	МатАн	ДифУр	ДисМат	ИИ	ФормГр	СисПрог	ЯзПрог	В целом
# слова	76 093	19 156	31 085	31 452	17 720	52 515	39 015	267 036
# уникальные термины	360	44	163	95	69	294	106	1 131
# обучающие пары	6 056	1 148	2 948	1 868	1 256	5 620	4 738	23 622
# вхождения терминов	657	854	1 181	250	329	2 516	2 040	7 827

Вторая часть рассматриваемого набора данных была создана для обучения модели последовательной разметки терминов в текстах. В текстах исходной коллекции семи предметных областей найдены вхождения терминов из известных для каждой области списков, и каждое найденное вхождение термина размечено на основе BIO-разметки (B − начальное слово-токен термина, I − внутренние слова, O − слова, не являющиеся частью термина), например: ...*пересечением*[B-term] *множеств*[I-term] *называется*[O] *множество*[B-term] *которое*[O]...

Всего было размечено 7 827 вхождений (терминоупотреблений), их количество по областям представлено в последней строке таблицы 1. Нейросетевая модель обучалась предсказывать одну из трех меток (B, I, O), согласно которым термины могут быть затем извлечены в список.

Программные модели распознавания терминов русского языка: обучение

Для экспериментов по распознаванию русскоязычных терминов на базе нейросетевых языковых моделей были выбраны две предобученные модели BERT, наиболее часто применяемые в настоящий момент для автоматического анализа русскоязычных текстов:

– ruBert-base-cased от проекта DeepPavlov [28] (далее ruBert-DeepPavlov) − многоязычный BERT [14], доработанный на русской википедии и новостных корпусах;

– ruBert-base от SberDevices (далее ruBert-Sber) [29] − BERT, исходно обученный для русского языка на текстах из русской википедии, новостей, книг, веб-сайтов и субтитров к фильмам.

Для дообучения этих моделей в обоих исследуемых подходах к распознаванию терминов (бинарная классификация, последовательная разметка) рассматривались три варианта разбиения созданного набора данных на подмножества для обучения, валидации и тестирования. В таблице 2 представлена информация об этих вариантах: составе каждого подмножества и количестве примеров в каждом из них. В первом варианте для обучения берутся данные из математики (условное название варианта − Математика), для валидации используются примеры из областей искусственного интеллекта и формальных грамматик, а для тестирования − данные из программирования. Второй вариант (Программирование) строился как противоположный первому (обучение на программировании, тестирование на математике). В третьем варианте области математики и программирования перемешаны. Рассмотрение таких вариантов позволило исследовать зависимость качества распознавания терминов от выбора предметной области для обучения.

При разбиении на подмножества для обучения, валидации и тестирования, кроме близости объединяемых предметных областей, были также учтены стандартные в машинном обучении пропорции данных в этих подмножествах. Отметим, что во всех вариантах разбиения доля общих терминов для обучающего и тестового подмножеств сравнительно мала, менее 4−7 %. Общие термины не исключались, поскольку это неизбежное явление для близких областей, которое следовало учитывать в экспериментах.

В предварительных экспериментах по обучению моделей была выполнена настройка различных гиперпараметров моделей. Для достижения наилучших значений F1-меры для бинарной классификации оказалось достаточно трех эпох обучения с оптимизатором AdamW и скоростью обучения 5e-6 , а для по-

Таблица 2

Варианты разбиения набора данных

Table 2

Dataset splitting options

Разбиение набора данных для обучения	Обучение		Валидация		Тестирование	
	Области	#	Области	#	Области	#
1 Математика	МатАн, ДифУр, ДисМат	10 152	ИИ ФормГр	3 134	СисПрог ЯзПрог	10 346
2 Программирование	СисПрог, ЯзПрог, ИИ, ФормГрам	13 470	ДифУр ДисМат	4 096	МатАн	6 056
3 Смешение областей	ДифУр, ДисМат, ИИ, ФормГр, ЯзПрог	11 946	СисПрог	5 620	МатАн	6 056

следовательной разметки понадобилось 10 эпох со скоростью обучения 5e-5 (задача последовательной разметки оказалась сложнее для обучения).

В итоге было обучено по шесть BERT-моделей для каждого из двух рассмотренных подходов:

– бинарный классификатор термин/нетермин, входными данными для которого служит пара <предложение + словосочетание из него>;

– классификатор токенов при последовательной разметке текста, входом которого является неразмеченный текст, в котором размечаются токены-слова, входящие в термины.

Программные модели распознавания терминов русского языка: результаты

Для всех обученных моделей качество классификации (то есть распознавания терминов) оценивалось по показателям точности, полноты и F1-меры, результаты представлены в таблице 3. Все модели демонстрируют приемлемые результаты по F1-мере от 49.1 до 73.3 %, но бинарные классификаторы для всех вариантов обучения (разбиения) и для всех моделей имеют ощутимо лучшие результаты – от 66.9 до 73.3 % F1-меры.

Сравнение по предобученным моделям показывает, что модели *ruBert/Sber* в среднем показывают лучшие результаты по F1-мере, но разрыв с другой моделью не очень большой: 53.5 % против 51.6 % для последовательной разметки и 72.0 % против 68.9 % для бинарной классификации.

Точность распознавания терминов (P) у всех моделей оказалась довольно высокой: до 84.5 % для бинарного классификатора и 79.9 % для последовательной разметки, но полнота (R) устойчиво ниже точности, хотя для бинарной классификации приемлема: в интервале от 55.6 до 70.5 %.

Сравнение двух подходов показало, что модели последовательной разметки немного проигрывают моделям бинарной классификации в точности (хотя на варианте Программирование модель *ruBert-Sber* оказалась лучше), но существенно проигрывает в полноте, причем на всех разбиениях (в среднем 40.7 % против 63.5 %).

Следует отметить, что в каждом из подходов качество моделей для разных разбиений набора данных по областям отличается незначительно, означая тем самым, что смена области от обучения к тестированию проходит без существенного падения качества − это может быть применено для распознавания терминов в новых (но близких) областях, для которых нет еще размеченных данных для обучения.

Кроме сопоставления эффективности подходов и качества моделей при разных разбиениях набора данных, было проведено сравнение с результатами традиционного статистического подхода на тех же исходных текстах по математике и программированию, по которым строился обучающий набор данных. Сравниваемый статистический метод опирался

– на грамматические образцы именных словосочетаний русского языка и шаблоны типичных контекстов употреблений терминов для выделения терминов-кандидатов;

Таблица 3

Оценки качества распознавания терминов

Table 3

Quality assessments of term recognition

Модель	Разбиение набора данных	Подход					
		Последовательная разметка			Бинарная классификация		
		P, %	**R, %**	**F1, %**	**P, %**	**R, %**	**F1, %**
ruBert-DeepPavlov	Математика	76.1	40.4	52.8	83.9	55.6	66.9
	Программирование	78.9	37.4	50.8	80.3	58.1	67.4
	Смешение	63.5	42.8	51.1	75.7	69.6	72.5
В среднем	по разбиениям	72.8	40.2	51.6	80.0	61.1	68.9
ruBert-Sber	Математика	75.9	43.3	55.1	**84.5**	63.9	72.8
	Программирование	**79.9**	35.4	49.1	77.6	63.2	70.0
	Смешение	74.0	**44.8**	**55.8**	76.3	**70.5**	**73.3**
В среднем	по разбиениям	76.6	41.2	53.4	79.5	65.9	**72.0**
Среднее по моделям		74.7	40.7	52.5	79.7	63.5	**70.5**

– на фильтрацию терминов-кандидатов (исключались стоп-слова и некоторые словосочетания с ними);

– на применение статистической меры терминологичности C-value для упорядочения терминов-кандидатов.

Результирующие оценки подходов сведены в таблицу 4, которая показывает, что наилучшим является подход с бинарной классификацией терминов, а подход с последовательной разметкой имеет примерно такое же качество извлечения терминов, что и статистический метод.

Таблица 4

Сравнение эффективности подходов к распознаванию терминов

Table 4

Comparing effectiveness of term recognition approaches

Подход	*P*, %	*R*, %	*F*1, %
Последовательная разметка	74.7	40.7	52.5
Бинарная классификация	79.7	63.5	**70.5**
Статистический метод	52.0	56.0	52.0

Результаты моделей распознавания терминов для русского языка, созданных на базе предобученных нейросетевых языковых моделей, показали, что качество бинарного BERT-классификатора сравнимо или немного лучше аналогов для других языков (английского, французского, словенского) [15, 16, 18]. Что же касается классификатора для последовательной разметки, то обученные модели, представленные авторами, имеют лучшие результаты, чем их аналоги в работе [18], сравнимы с аналогами из [16], но хуже, чем в [17]. Низкие результаты моделей, возможно, объясняются меньшим размером набора данных для обучения.

Проведенный вручную анализ ошибок распознавания терминов, допущенных наилучшими моделями, показал, что не были распознаны как термины некоторые словосочетания (например, *императивная парадигма*), которые в действительности таковыми являются (это свидетельствует об ошибках в разметке набора данных, которые в дальнейшем могут быть устранены). Более сложными оказались ложноотрицательные (например, *уровень списка*) и ложноположительные случаи (например, *набор функций*), когда словосочетания содержали слова общей лексики (*уровень* и *набор*). Заметим, однако, что использование таких слов в качестве элементов термина затрудняет распознавание терминов даже людьми.

Заключение

В работе охарактеризованы известные подходы к задаче автоматического распознавания и извлечения терминов из текстов. Проведено исследование современного подхода на базе нейросетевой языковой модели BERT применительно к русскому языку и с учетом перехода на близкую предметную область. Оценка качества распознавания терминов разработанными программными моделями показала, что бинарная классификация термин/нетермин при наличии контекстного предложения превосходит метод на основе последовательной разметки терминов в тексте, а также статистический метод. Для обучения моделей и проведения исследования был построен репрезентативный открыто доступный набор размеченных данных, охватывающий термины из семи научных областей математики и программирования и применимый для дальнейшего развития рассмотренной задачи.

Список литературы

1. Vivaldi J., Rodrigues H. Evaluation of terms and term extraction systems: A practical approach. Terminology, 2007, vol. 13, no. 2, pp. 225–248. doi: 10.1075/term.13.2.06viv.

2. Korkontzelos I., Ananiadou S. Term extraction. In: The oxford handbook of computational linguistics, 2014, pp. 991–1012. doi: 10.1093/oxfordhb/9780199573691.013.004.

3. Loukachevitch N., Nokel M. An experimental study of term extraction for real information-retrieval thesauri. Proc. TIA-2013, 2013, pp. 69–76.

4. Соловьев С.Ю. Схема и формула глоссария // КИИ-2008: тр. конф. 2008. Т. 2. С. 157–164.

5. Мальковский М.Г., Соловьев С.Ю. От терминологических сетей к толковым словарям // OSTIS-2015: мат. Междунар. науч.-технич. конф. 2015. С. 281–284.

6. Bolshakova E.I., Ivanov K.M. Automating hierarchical subject index construction for scientific documents. In: LNAI. Proc. RCAI-2020, 2020, vol. 12412, pp. 201–214. doi: 10.1007/978-3-030-59535-7_14.

7. Arcan M., Turchi M., Tonelli S., Buitelaar P. Leveraging bilingual terminology to improve machine translation in a CAT environment. Natural Language Eng., 2017, vol. 23, no. 5, pp. 763–788. doi: 10.1017/S1351324917000195.

8. Pazienza M.T., Pennacchiotti M., Zanzotto F.M. Terminology extraction: An analysis of linguistic and statistical approaches. In: STUDFUZZ. Proc. Knowledge Mining, 2005, vol. 185, pp. 255–279. doi: 10.1007/3-540-32394-5_20.

9. Zhang Z., Iria J., Brewster C., Ciravegna F. A comparative evaluation of term recognition algorithms. Proc. Int. Conf. LREC'08, 2008, pp. 2108–2111.

10. Zhang Z., Gao J., Ciravegna F. SemRe-Rank: Improving automatic term extraction by incorporating semantic relatedness with personalised PageRank. ACM TKDD, 2018, vol. 12, no. 5, art. 57, pp. 1–41. doi: 10.1145/3201408.

11. Yuan Y., Gao J., Zhang Y. Supervised learning for robust term extraction. Proc. IALP, 2017, pp. 302–305. doi: 10.1109/IALP.2017.8300603.

12. Terryn A.R., Drouin P., Hoste V., Lefever E. Analysing the impact of supervised machine learning on automatic term extraction: HAMLET Vs TermoStat. Proc. Int. Conf. RANLP, 2019, pp. 1013–1022. doi: 10.26615/978-954-452-056-4_117.

13. Terryn A.R., Hoste V., Lefever E. Tagging terms in text: A supervised sequential labeling approach to automatic term extraction. Terminology, 2022, vol. 28, no. 1, pp. 157–189. doi: 10.1075/term.21010.rig.

14. Devlin J., Chang M., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. Proc. Conf. of the North, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

15. Hazem A., Bouhandi M., Boudin F., Daille B. TermEval 2020: TALN-LS2N system for automatic term extraction. Proc. Int. Workshop on Computational Terminology, 2020, pp. 95–100.

16. Lang C., Wachowiak L., Heinisch B., Gromann D. Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains. Proc. ACL-IJCNLP, 2021, pp. 3607–3620. doi: 10.18653/v1/2021.findings-acl.316.

17. Tran H.T.H., Martinc M., Repar A., Doucet A., Pollak S. A transformer-based sequence-labeling approach to the slovenian cross-domain automatic term extraction. Proc. Slovenian Conf. on Language Technologies and Digital Humanities, 2022, pp. 196–204.

18. Hazem A., Bouhandi M., Boudin F., Daille B. Cross-lingual and cross-domain transfer learning for automatic term extraction from low resource data. Proc. LREC, 2022, pp. 648–662.

19. Frantzi K., Ananiadou S., Mima H. Automatic recognition of multi-word terms: The C-value/NC-value method. IJDL, 2000, vol. 3, no. 2, pp. 115–130. doi: 10.1007/s007999900023.

20. Šajatović A., Buljan M., Snajder J., Basic B.D. Evaluating automatic term extraction methods on individual documents. Proc. Joint Workshop MWE-WN, 2019, pp. 149–154. doi: 10.18653/v1/W19-5118.

21. Большакова Е.И., Иванов К.М. Извлечение терминов для построения предметного указателя учебно-научного текста // Компьютерная лингвистика и интеллектуальные технологии: мат. конф. «Диалог». 2018. № 17. С. 143–152 (на англ.).

22. Cram D., Daille B. TermSuite: Terminology extraction with term variant detection. Proc. 54th Annual Meeting ACL, 2016, pp. 13–18. doi: 10.18653/v1/P16-4003.

23. Marciniak M., Mykowiecka A., Rychlik P. TermoPL – a flexible tool for terminology extraction. Proc. 10th Int. Conf. LREC'16, 2016, pp. 2278–2284.

24. Zhang Z., Gao J., Ciravegna F. Jate 2.0: Java automatic term extraction with apache solr. Proc. 10th Int. Conf. LREC'16, 2016, pp. 2262–2269.

25. Astrakhantsev N. ATR4S: toolkit with state-of-the-art automatic terms recognition methods in Scala. LRE, 2018, vol. 52, pp. 853–872. doi: 10.1007/s10579-017-9409-4.

26. Машонский И.Д., Большакова Е.И. Инструментальные средства извлечения терминов из текстов: разработка компонентов для русского языка // Программные системы и инструменты: Тематический сборник. 2020. № 20. С. 94–105.

27. Terryn A.R., Hoste V., Drouin P., Lefever E. TermEval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (ACTER) dataset. Proc. Int. Workshop COMPUTERM, 2020, pp. 85–94.

28. Куратов Ю., Архипов М. Адаптация глубоких двунаправленных многоязычных моделей на основе архитектуры transformer для русского языка // Компьютерная лингвистика и интеллектуальные технологии: мат. конф. «Диалог». 2019. № 18. С. 310–316 (на англ.).

29. Zmitrovich D., Abramov A., Kalmykov A., Kadulin V. et al. A family of pretrained transformer language models for Russian. Proc. LREC-COLING, 2024, pp. 507–524.

Software & Systems doi: 10.15827/0236-235X.149.005-016 2025, 38(1), pp. 5–16

Methods and means of term extraction from texts for terminological tasks

Elena I. Bolshakova [1]✉, Vladislav V. Semak [1]

[1] Lomonosov Moscow State University, Moscow, 119991, Russian Federation

For citation

Bolshakova, E.I., Semak, V.V. (2025) 'Methods and means of term extraction from texts for terminological tasks', *Software & Systems*, 38(1), pp. 5–16 (in Russ.). doi: 10.15827/0236-235X.149.005-016

Article info

Received: 18.08.2024 After revision: 14.09.2024 Accepted: 29.09.2024

Abstract. The paper describes the current state in the field of automatic term extraction from specialized natural language texts, including scientific and technical documents. Practical applications of methods and tools for extracting terms from texts include creation of terminological dictionaries, thesauri, and glossaries of problem oriented domains, as well as extraction of keywords and construction of subject indexes for specialized documents. The paper overviews approaches to automatic recognition and extraction of terminological words and phrases, which cover traditional statistical methods and methods based on machine learning by using term features or with modern neural network transformer-based language models. A comparison of the approaches is presented, including quality assessments for term recognition and term extraction. The most well-known software tools for automating term extraction within the statistical approach and learning by features are indicated. Authors' studies on term recognition based on neural network language models are described, being applied to Russian scientific texts on mathematics and programming. The data set with terminological annotations created for training term recognition models is briefly characterized, the dataset covers the data from seven related domains. The term recognition models were developed on the basis of pre-trained neural network model BERT, with its additional training (fine-tuning) in two ways: as binary classifier of candidate terms (previously extracted from texts) and as classifier for sequential labeling words in texts. For the developed models, the quality of term recognition is experimentally evaluated, and a comparison with the statistical approach was carried out. The best quality is demonstrated by binary classification models, significantly surpassing the other considered approaches. The experiments also show the applicability of the trained models to texts in close scientific domains.

Keywords: automatic processing of specialized texts, terminological words and phrases, automatic term extraction, machine learning for term recognition, BERT-based term extraction

References

1. Vivaldi, J., Rodrigues, H. (2007) ‘Evaluation of terms and term extraction systems: A practical approach’, *Terminology*, 13(2), pp. 225–248. doi: 10.1075/term.13.2.06viv.

2. Korkontzelos, I., Ananiadou, S. (2014) ‘Term extraction’, in *The Oxford Handbook of Computational Linguistics*, pp. 991–1012. doi: 10.1093/oxfordhb/9780199573691.013.004.

3. Loukachevitch, N., Nokel, M. (2013), ‘An experimental study of term extraction for real information-retrieval thesauri’, *Proc. TIA-2013*, pp. 69–76.

4. Solovev, S.Yu. (2008) ‘Glossary diagram and formula’, *Proc. CAI-2008*, 2, pp. 157–164 (in Russ.).

5. Malkovsky, M.G., Soloviev, S.Yu. (2015) ‘From terminological networks to the explanatory dictionaries’, *Proc. Int. Sci. Tech. Conf. OSTIS-2015*, pp. 281–284 (in Russ.).

6. Bolshakova, E.I., Ivanov, K.M. (2020) ‘Automating hierarchical subject index construction for scientific documents’, in *LNAI. Proc. RCAI-2020*, 12412, pp. 201–214. doi: 10.1007/978-3-030-59535-7_14.

7. Arcan, M., Turchi, M., Tonelli, S., Buitelaar, P. (2017) ‘Leveraging bilingual terminology to improve machine translation in a CAT environment’, *Natural Language Eng*., 23(5), pp. 763–788. doi: 10.1017/S1351324917000195.

8. Pazienza, M.T., Pennacchiotti, M., Zanzotto, F.M. (2005) ‘Terminology extraction: An analysis of linguistic and statistical approaches’, In: *STUDFUZZ. Proc. Knowledge Mining*, 185, pp. 255–279. doi: 10.1007/3-540-32394-5_20.

9. Zhang, Z., Ira, J., Brewster, C., Ciravegna, F. (2008) ‘A comparative evaluation of term recognition algorithms’, *Proc. Int. Conf. LREC’08*, pp. 2108–2111.

10. Zhang, Z., Gao, J., Ciravegna, F. (2018) ‘SemRe-Rank: Improving automatic term extraction by incorporating semantic relatedness with personalised PageRank’, *ACM TKDD*, 12(5), art. 57, pp. 1–41. doi: 10.1145/3201408.

11. Yuan, Y., Gao, J., Zhang, Y. (2017) ‘Supervised learning for robust term extraction’, *Proc. IALP*, pp. 302–305. doi: 10.1109/IALP.2017.8300603.

12. Terryn, A.R., Drouin, P., Hoste, V., Lefever, E. (2019) ‘Analysing the impact of supervised machine learning on automatic term extraction: HAMLET Vs TermoStat’, *Proc. Int. Conf. RANLP*, pp. 1013–1022. doi: 10.26615/978-954-452-056-4_117.

13. Terryn, A.R., Hoste, V., Lefever, E. (2022) ‘Tagging terms in text: A supervised sequential labelling approach to automatic term extraction’, *Terminology*, 28(1), pp. 157–189. doi: 10.1075/term.21010.rig.

14. Devlin, J., Chang, M., Lee, K., Toutanova, K. (2019) ‘BERT: Pre-training of deep bidirectional transformers for language understanding’, *Proc. Conf. of the North*, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

15. Hazem, A., Bouhandi, M., Boudin, F., Daille, B. (2020) ‘TermEval 2020: TALN-LS2N system for automatic term extraction’, *Proc. Int. Workshop on Computational Terminology*, pp. 95–100.

16. Lang, C., Wachowiak, L., Heinisch, B., Gromann, D. (2021) ‘Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains’, *Proc. ACL-IJCNLP*, 2021, pp. 3607–3620. doi: 10.18653/v1/2021.findings-acl.316.

17. Tran, H.T.H., Martinc, M., Repar, A., Doucet, A., Pollak, S. (2022) ‘A transformer-based sequence-labeling approach to the slovenian cross-domain automatic term extraction’, *Proc. Slovenian Conf. on Language Technologies and Digital Humanities*, pp. 196–204.

18. Hazem, A., Bouhandi, M., Boudin, F., Daille, B. (2022) ‘Cross-lingual and cross-domain transfer learning for automatic term extraction from low resource data’, *Proc. LREC*, pp. 648–662.

19. Frantzi, K., Ananiadou, S., Mima, H. (2000) ‘Automatic recognition of multi-word terms: The C-value/NC-value method’, *IJDL*, 3(2), pp. 115–130. doi: 10.1007/s007999900023.

20. Šajatović, A., Buljan, M., Snajder, J., Basic, B.D. (2019) ‘Evaluating automatic term extraction methods on individual documents’, *Proc. Joint Workshop MWE-WN*, pp. 149–154. doi: 10.18653/v1/W19-5118.

21. Bolshakova, E.I., Ivanov, K.M. (2018) ‘Term extraction for constructing subject index of educational scientific text’, *Computational Linguistics and Intellectual Technologies: Proc. Conf. “Dialogue”*, pp. 143–152.
22. Cram, D., Daille, B. (2016) ‘TermSuite: Terminology extraction with term variant detection’, *Proc. 54th Annual Meeting ACL*, pp. 13–18. doi: 10.18653/v1/P16-4003.
23. Marciniak, M., Mykowiecka, A., Rychlik, P. (2016) ‘TermoPL – a flexible tool for terminology extraction’, *Proc. 10th Int. Conf. LREC’16*, pp. 2278–2284.
24. Zhang, Z., Gao, J., Ciravegna, F. (2016) ‘Jate 2.0: Java automatic term extraction with apache solr’, *Proc. 10th Int. Conf. LREC’16*, pp. 2262–2269.
25. Astrakhantsev, N. (2018) ‘ATR4S: toolkit with state-of-the-art automatic terms recognition methods in Scala’, *LRE*, 52, pp. 853–872. doi: 10.1007/s10579-017-9409-4.
26. Mashonsky, I.D., Bolshakova, E.I. (2020) ‘Tools for terminology extraction from texts: development of components for russian language’, *Software Systems and Tools*, (20), pp. 94–105 (in Russ.).
27. Terryn, A.R., Hoste, V., Drouin, P., Lefever, E. (2020) ‘Termeval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (ACTER) dataset’, *Proc. Int. Workshop COMPUTERM*, pp. 85–94.
28. Kuratov, Yu., Arkhipov, M. (2019) ‘Adaptation of deep bidirectional multilingual transformers for Russian language’, *Computational Linguistics and Intellectual Technologies: Proc. Conf. “Dialogue”*, (18), pp. 310–316.
29. Zmitrovich, D., Abramov, A., Kalmykov, A., Kadulin, V. et al. (2024) ‘A family of pretrained transformer language models for Russian’, *Proc. LREC-COLING*, pp. 507–524.

Авторы
Большакова Елена Игоревна [1],
к.ф.-м.н., доцент,
bolsh@cs.msu.ru
Семак Владислав Викторович [1],
аспирант, vlad.semakk@gmail.com

[1] Московский государственный университет
им. М.В. Ломоносова, г. Москва, 119991, Россия

Authors
Elena I. Bolshakova [1],
Cand. of Sci. (Physics and Mathematics),
Associate Professor, bolsh@cs.msu.ru
Vladislav V. Semak [1],
Postgraduate Student, vlad.semakk@gmail.com

[1] Lomonosov Moscow State University,
Moscow, 119991, Russian Federation