УДК 004.912 **DOI** 10.15622/ia.23.4.7

В.К. Пимешков, М.Л. Никонорова, М.Г. Шишаев

КОМБИНИРОВАННЫЙ МЕТОД ИЗВЛЕЧЕНИЯ ТЕРМИНОВ ДЛЯ ЗАДАЧИ МОНИТОРИНГА ТЕМАТИЧЕСКИХ ОБСУЖДЕНИЙ В СОЦИАЛЬНЫХ МЕДИА

Пимешков В.К., Никонорова М.Л., Шишаев М.Г. **Комбинированный метод извлечения терминов для задачи мониторинга тематических обсуждений в социальных медиа.**

Аннотация. Извлечение терминов является важным этапом автоматизированного построения систем знаний на основе естественно-языковых текстов, поскольку обеспечивает формирование базовой системы понятий, используемой затем в прикладных задачах интеллектуальной обработки информации. В статье рассмотрена проблема автоматизированного извлечения терминов из естественно-языковых текстов с целью их дальнейшего использования при построении формализованных систем знаний (онтологий, тезаурусов, графов знаний) в рамках задачи мониторинга тематических обсуждений в социальных медиа. Данная задача характеризуется необходимостью включения в формируемую систему знаний как понятий из нескольких различных предметных областей, так и некоторых общеупотребительных понятий, используемых аудиторией социальных медиа в рамках тематических обсуждений. Кроме того, формируемая система знаний является динамичной как с точки зрения состава охватываемых ею предметных областей, так и состава релевантных понятий, подлежащих включению в систему. Применение существующих классических методов извлечения терминов в данном случае затруднительно, поскольку они ориентированы на извлечение терминов в рамках одной предметной области. Исходя из этого, для решения рассматриваемой задачи предложен комбинированный метод, совмещающий в себе подходы на основе внешних источников знаний, инструментов NER и правил. Результаты проведенных экспериментов демонстрируют эффективность предложенной комбинации подходов к извлечению терминов для задачи мониторинга и анализа тематических обсуждений в социальных медиа. Разработанный метод значительно превосходит по точности существующие инструменты извлечения терминов. В качестве дальнейшего направления исследования рассмотрена возможность развития метода для решения задачи выделения вложенных терминов или сущностей.

Ключевые слова: интеллектуальный анализ текстов, извлечение терминов, социальные медиа, извлечение знаний.

1. Введение. Задача извлечения терминов (term extraction) представляет собой процесс автоматического выделения ключевых терминов или слов из текста или набора текстов с целью определения наиболее репрезентативных слов в контексте рассматриваемой прикладной задачи. В данной работе рассматривается специфический класс подобных прикладных задач, характеризуемый множественностью предметных областей, с одной стороны, и важностью учета общеупотребительной лексики – с другой. Примером такой задачи является мониторинг и анализ социально-политической ситуации на некоторой ограниченной территории (город, регион и т.п.). Использование для ее решения данных

социальных медиа сопряжено с проблемой извлечения структурированной информации из огромного объема постоянно генерируемых сырых данных, составляющих содержание тематических обсуждений в социальных медиа. Базовым компонентом такой информации являются термины, являющиеся лексическими обозначениями понятий, используемых участниками дискуссий. В то же время, специфика как лексических особенностей коммуникаций в социальных медиа в целом, так и рассматриваемой прикладной задачи, ограничивает возможность применения для извлечения терминов известных готовых технологий и методов.

Извлечение терминов является критически важным этапом для задачи извлечения знаний – концептов (единиц мысли) и связей между ними. Процесс извлечения знаний из текстов заключается в выделении всех релевантных (обозначающих значимые для последующего использования концепты) терминов и в дальнейшем определении прямых или опосредованных связей между ними. При этом релевантность, чаще всего, определяется принадлежностью термина некоторой предметной области. В свою очередь, предметная область задается языковым ресурсом – коллекцией текстов соответствующей тематики, словарем, тезаурусом и т.п.

Как правило, общение в социальных медиа хоть и представлено в виде текста, но имеет присущие живой разговорной речи особенности, такие как свернутость, спонтанность, экспрессивность, высокая доля лексики с разговорной окраской, обилие специфических разговорных конструкций и т.д. [1]. Помимо этого, для социальных медиа характерна еще одна особенность – мультипредметность: наличие/употребление терминов из многих предметных областей. При этом мы различаем предметные термины (специфичные для некоторой предметной области) и общеупотребительные термины, в равной степени широко используемые в коммуникациях любой тематики. Важной особенностью социальных медиа как источника релевантных терминов является отсутствие каких-либо ограничений (присущих профессиональной, предметной, коммуникации) в использовании терминологии. Как следствие, интересующие нас понятия могут обозначаться как предметными терминами, так и терминами общей лексики.

Таким образом, в отличие от классической задачи извлечения терминов, где требуется извлечение ограниченного набора релевантных терминов в рамках одной предметной области, в задаче мониторинга тематических обсуждений нас интересуют не только предметные, но и общеупотребительные термины. Кроме того,

предметные термины в нашем случае могут принадлежать нескольким различным предметным областям, лексический состав которых может изменяться, если мониторинг осуществляется достаточно длительное время. В таких условиях вполне ожидаемо, что готовые словари или тезаурусы с подобными терминами в открытом доступе отсутствуют. Данная особенность рассматриваемой в работе задачи иллюстрируется рисунком 1.

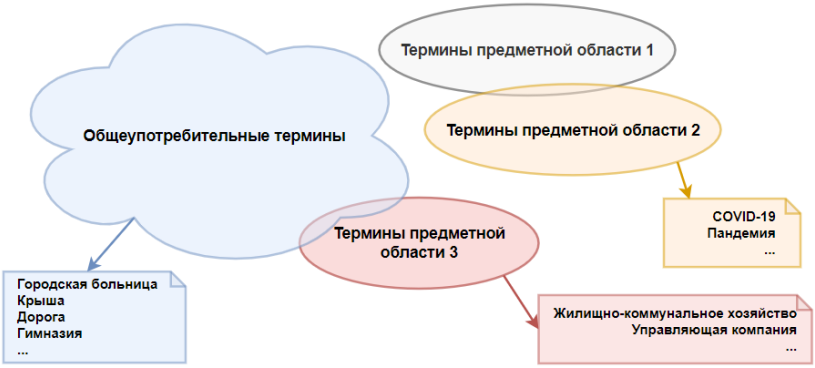

Рис. 1. Структура лексики коммуникаций в рамках рассматриваемой прикладной задачи

В результате, применение только классических методов извлечения терминов не покрывает все интересующие нас термины и, следовательно, не обеспечивает достаточной полноты в рамках решаемой задачи. Это обусловливает необходимость разработки специального метода, учитывающего динамичность тематических обсуждений, выражающуюся в изменении набора интересующих нас предметных областей или набора интересующих нас терминов в рамках одной предметной области. В данной работе предложен подобный метод, основанный на комбинировании различных подходов к извлечению терминов, и проведена оценка его эффективности на рассматриваемом классе прикладных задач. Особенностью метода является учет структуры коммуникации и пользовательских реакций при принятии решения о включении некоторого термина в число значимых. В процессе разработки и валидации предложенного метода использовались данные, аккумулированные ранее в рамках мониторинга социально-политической обстановки в Кировско-Апатитском регионе Мурманской области: размеченный датасет, сформированный из сообщений сети ВКонтакте, а также коллекция соответствующих региональных нормативно-правовых документов.

2. Методы извлечения терминов для создания формализованных систем знаний

2.1. Краткая характеристика существующих («классических») методов извлечения терминов. Классические методы извлечения терминов можно разделить на три категории: методы на основе правил, методы, использующие словари или внешние источники знаний, а также статистические методы. Вне зависимости от метода задача заключается в поиске *значимых (релевантных)* терминов – терминов, важных в контексте рассматриваемой прикладной задачи.

Методы на основе правил используют лексические или морфосинтаксические шаблоны для извлечения значимых в рамках задачи терминов. Значимость термина в данной категории методов определяется правилами, задаваемыми экспертно или с помощью автоматизированных методов. Формально данный метод извлечения можно представить следующим образом:

Дано:

$D = \{d_1, \dots, d_N\}$ – множество (коллекция) документов различных тематик, при этом $D = D^S \cup D^G$, где D^S – множество документов предметной области, D^G – множество документов общей тематики;

$P = \{p_1, \dots, p_K\}$ – множество правил извлечения терминов, где:

$p_i: D \to B(T)$, $B(\cdot)$ – булеан множества, T – множество всевозможных терминов.

Найти:

$T(D) = \bigcup_i T_i$ – множество соответствующих правилам терминов из коллекции документов D, где $T_i = \bigcup_{j=1}^{N} p_i(d_j)$ – множество терминов, соответствующих правилу p_i.

Например, в [2] используются морфологические правила, определенные на основе анализа предметных текстов, для извлечения типизированных структур данных. В [3] используют лингвистические шаблоны, дополненные контекстными правилами, принимающие во внимание фразы с предлогами и причастиями, для повышения эффективности извлечения терминов. В рамках такого метода сначала извлекаются структуры, содержащие определенные предлоги. Полученные кандидаты фильтруются от терминов неверной структурной формы (т.е. ошибки токенизации и частеречевой разметки) и от семантически бедных терминов (грамматически правильных, но не значимых для прикладных задач) с помощью набора правил. Авторы также уточняют частеречевую разметку для особых форм глаголов, что позволяет принять во внимание больший спектр потенциальных терминов. Результаты применения такого

метода могут использоваться для решения таких задач, как создание и пополнение онтологий и терминологий.

Методы, основанные на внешних источниках знаний, используют существующие терминологические ресурсы (тезаурусы, словари, онтологии) с целью поиска вхождений, как правило, предметных терминов в тексте. Значимость в данном случае исчерпывающим образом определяется внешним источником (те и только те термины являются значимыми, которые в некотором смысле схожи с входящими в источник). Формализация данной категории методов выглядит следующим образом:

Дано:

D – множество документов различных тематик;

$T^S = \{t_1^S, \dots, t_M^S\}$ – множество терминов источника знаний S, определяющего предметную область;

$sim : T \times T \to \{0,1\}$ – некоторая функция схожести термина документа с термином источника знаний, $sim(t_1, t_2) = 1$, если термины t_1, t_2 схожи и $sim(t_1, t_2) = 0$ в противном случае.

Найти:

$T^{S+} = \{t_i^{S+} | \forall t_i^{S+} : \exists t^S \in T^S : f(t_i^{S+}, t^S) = 1\}$ – множество терминов, схожих в смысле функции f с содержащимися во внешнем источнике.

Например, предложенный в работе [4] подход к извлечению биологических терминов основан на приближенном поиске по словарю, который позволяет извлекать из текста не только точные совпадения терминов, но и значимые слова этих терминов. Значимость слова определяется с помощью соответствующей меры значимости, обученной на основе словаря. В системе извлечения информации [5] в качестве исходного источника знаний выступает онтология, которая в процессе преобразуется в словарь, состоящий из категорий и соответствующих им сущностей. Затем используется анализатор, содержащий сформированный на основе данного словаря набор правил, для поиска в тексте совпадений именных групп, представляющих важные медицинские сущности.

Стоит отметить, что задача пополнения онтологии/тезауруса может рассматриваться как частный случай использования методов извлечения терминов на основе внешних источников. При решении данной задачи на входе также имеется внешний ресурс и текст, а также некоторая функция, возвращающая степень схожести термина-кандидата и термина, присутствующего в исходном ресурсе. Функция схожести может быть реализована различными способами: схожесть контекста [6], схожесть векторного представления в рамках некоторой языковой модели [7] и т.д. Частный случай – когда функция схожести

бинарная: возвращает 1 для полностью идентичных терминов и 0 для прочих, что соответствует вышеизложенному определению данной категории методов.

Статистические методы используют анализ большого количества наблюдений для идентификации терминов. Такие методы редко применяются без какого-либо этапа синтаксического или лексического анализа ввиду того, что это может приводить к большому объему нежелательной терминологии [8]. В статистических методах признак значимости термина определяется частотностью термина в наборе предметных документов. Значимость при этом трактуется как «предметность» – соответствие термина интересующей нас предметной области. Формальный вид данной категории методов следующий:

Дано:

$D^S = \{ d_1^S, \ldots, d_N^S \}$ – множество документов предметной области;

$D^G = \{ d_1^G, \ldots, d_N^G \}$ – множество документов общей тематики;

$m: B(D) \times T$ – статистическая мера, характеризующая частоту встречаемости термина из T во множестве документов из $B(D)$, где $B(\cdot)$ – булеан множества.

Найти:

$T^S: t^S \in T^S \leftrightarrow m(D^S, t^S) > m(D^G, t^S)$ – множество предметных терминов.

Зачастую статистические методы объединяются с другими для достижения лучших результатов. Например, в [9] используются коллокации (частный случай N-грамм), дополненные частеречной и статистической фильтрацией для извлечения терминов-кандидатов из предметных текстов на немецком языке. Данный метод включает четыре этапа. На первом этапе выполняется предварительная обработка текста, включающая разбиение текста на предложения и другие структурные составляющие, разбиение текста на токены и определение их частей речи. На втором этапе извлекаются однословные термины с помощью существующих процедур, основанных на частотных характеристиках, морфологии и списках морфем. На третьем этапе на основе грамматических свойств извлекаются коллокационные кандидаты-словосочетания. И четвертым этапом производится фильтрация словосочетаний-кандидатов, удаляются тривиальные словосочетания (содержащие определенные общие глаголы), а также общеупотребительные словосочетания. Термины, извлеченные таким методом, могут использоваться для создания и пополнения предметных терминологий.

Кроме того, для извлечения терминов могут использоваться инструменты извлечения именованных сущностей (Named Entity Recognition, NER), нацеленные на извлечение и классификацию именованных сущностей, которые могут рассматриваться как особые типы терминов. В классическом NER выделяют такие типы, как названия организаций, локаций и имена людей. Но NER находит применение и в конкретных предметных областях. Например, в биомедицине выделяют названия генов, белков и связанных с ними биологические или генетические термины, а также наименования болезней, лекарств и др. NER может быть реализован с помощью правил, словарей или машинного обучения. Например, в работе [10] предлагается метод извлечения и классификации биомедицинских терминов на основе NER без учителя. Для извлечения терминов-кандидатов авторы используют выделение именных групп с фильтрацией по TF-IDF (Term Frequency – Inverse Document Frequency), а для классификации этих кандидатов применяется мера схожести с шаблонами классов, разработанными авторами на основе внешних источников.

Помимо рассмотренных выше методов, распространенной практикой является использование гибридных методов, комбинирующих различные подходы. Наиболее популярной является комбинация лингвистических и статистических методов в том или ином виде. К примеру, в [11] используют решающие деревья для отбора и ранжирования терминов-кандидатов на основе трех различных наборов признаков. Первый набор включает в себя структурные признаки, полученные с помощью системы GROBID [12] (библиотека машинного обучения для извлечения, анализа и реструктуризации необработанных документов). Второй набор состоит из фразеологических и информативных (насколько термин отражает смысл документа в рамках набора документов) признаков, полученных различными статистическими методами. И третий набор содержит лексические и семантические признаки, полученные с помощью таких внешних источников знаний, как GRISP [13] и Википедия.

2.2. Комбинированный метод извлечения терминов для задач анализа тематических обсуждений. Предлагаемый нами комбинированный метод извлечения терминов использует комбинацию вышеупомянутых категорий методов. В нашем случае множество значимых терминов не ограничивается предметными, но также должно включать некоторое подмножество общеупотребительных терминов, ассоциированных с предметными.

Соответственно, значимость термина определяется источником знаний, определяющим множество предметных терминов, и некоторой метрикой ассоциированности общеупотребительных терминов с элементами этого множества. Для расчета последней в рамках метода предлагается использовать количественные показатели пользовательской реакции и структуру коммуникации в рамках тематического обсуждения. Формально предлагаемый метод можно представить следующим образом:

Дано:

D – множество документов; документ представляет собой пост (сообщение) или комментарий;

$l(d)$ – количество лайков документа d;

$r(d)$ – количество репостов документа d;

$c(d)$ – количество комментариев документа d;

$v(d)$ – количество просмотров документа d;

T^S – множество терминов базового источника знаний;

P – множество правил извлечения терминов;

M – языковая модель для извлечения именованных сущностей;

$ner: M \times D \to B(T)$ – функция извлечения именованных сущностей из документов с помощью модели M.

$L \subseteq D \times D$ – асимметричное транзитивное отношение «является откликом», определяющее на множестве документов структуру коммуникации. $d_1 L d_2$ означает, что документ (пост или комментарий) d_2 является откликом на документ d_1. Структура коммуникации состоит из веток обсуждения, представляющих собой последовательности Br документов из D ($Br \subseteq D$), удовлетворяющие условию:

$$Br = \{d_1, \dots, d_N\}: \forall i < j, d_i L d_j.$$

Метод извлечения терминов включает следующие основные шаги:

1) Извлечение из коллекции документов D множества предметных терминов T^{S+} на основе внешнего источника знаний T^S и функции схожести sim.

2) Извлечение из коллекции документов D множества терминов T^P, удовлетворяющих заданному набору правил P.

3) Извлечение из коллекции документов D множества именованных сущностей T^M с помощью функции ner.

4) Выделение из множества $T^P \cup T^M$ множества терминов T^A, ассоциированных с терминами из $T^S \cup T^{S+}$.

5) Результат работы метода есть объединение множеств извлеченных предметных терминов T^{S+} и значимых общеупотребительных терминов T^{A}.

Отбор терминов для включения в множество T^{A} осуществляется на основании метрики ассоциированности, которая зависит от контекста, частоты употребления термина s и интенсивности пользовательских реакций на документы-источники s. Метрика может быть представлена функционалом вида:

$$assc(t,s) = F\big(ct(t,s), q(s), fd(s)\big): t \in T^{S} \cup T^{S+}, s \in T\,,$$

где
$ct: T \times T \to \mathbb{R}$ – функция оценки контекста, удовлетворяющая следующим условиям:

$$ct(t_1,t_2) > 0, t_1 \in d_i, t_2 \in d_j \leftrightarrow \exists Br: d_i, d_j \in Br;$$

$$ct(t_1,t_2) > ct(t_1,t_3), t_1 \in d_i, t_2 \in d_j, t_3 \in d_k \leftrightarrow$$
$$\leftrightarrow \exists Br: d_i, d_j, d_k \in Br \wedge |i-j| < |i-k|;$$

$$ct(t_1,t_2) = 0, t_1 \in d_i, t_2 \in d_j \leftrightarrow \nexists Br: d_i, d_j \in Br.$$

$q(s)$ – частота встречаемости термина s;
$fd(s) = \sum_{i=1}^{n} fd(d_i) : s \in d_i$ – суммарная оценка реакции на термин s;
$fd(d_i) = f(l(d_i), r(d_i), c(d_i), v(d_i))$ – обобщенная оценка интенсивности реакции на документ d_i.

Следует отметить, что каждый документ может быть как источником некоторого инфоповода, так и откликом на него. При этом отклики могут вызывать свои собственные отклики, и в таком случае первые могут рассматриваться как источники. Вместе с тем адресат/инициатор отклика не всегда выражен в явном виде (например, комментарий или репост), т.к. вполне возможно появление отклика на какую-то популярную публикацию или событие без явной привязки к оному.

Сравнительная характеристика рассмотренных методов извлечения терминов представлена в таблице 1.

Таблица 1. Сравнение методов извлечения терминов

Метод	**Используемые исходные данные**	**Ожидаемый результат и ключевая особенность**
Метод на основе правил	$D = \{d_1, \ldots, d_N\}$ – множество (коллекция) документов различных тематик; $P = \{p_1, \ldots, p_K\}$ – множество правил извлечения терминов, где $p_i: D \to B(T)$, T – множество всевозможных терминов.	$T(D) = \bigcup_i T_i$ – множество соответствующих правилам терминов из коллекции документов D, где $T_i = \bigcup_{j=1}^{N} p_i(d_j)$ – множество терминов, соответствующих правилу p_i. Термины не различаются по признаку «предметности».
Метод на основе внешних источников	D – множество документов различных тематик; $T^S = \{t_1^S, \ldots, t_M^S\}$ – множество терминов источника знаний S, определяющего предметную область; $sim: T \times T \to \{0,1\}$ – функция схожести терминов.	$T^{S+} = \{t_i^{S+} \mid \forall t_i^{S+}: \exists t^S \in T^S: f(t_i^{S+}, t^S) = 1\}$ – множество терминов, схожих в смысле функции f с содержащимися во внешнем источнике. «Предметность» термина определяется внешним источником.
Статистический метод	D^S – множество документов предметной области; D^G – множество документов общей тематики; $m: B(D) \times T$ – статистическая мера, характеризующая частоту встречаемости термина в документах.	$T^S: t^S \in T^S \leftrightarrow m(D^S, t^S) > m(D^G, t^S)$ – множество предметных терминов. «Предметность» термина определяется соотношением предметной и обще-тематической коллекций документов.
Предлагаемый комбинированный метод	D – множество документов; $l(d), r(d), c(d), v(d)$ – количественные индикаторы пользовательской реакции на документ d; T^S – множество терминов базового источника знаний; P – множество правил извлечения терминов; M – языковая модель для извлечения именованных сущностей; $ner: M \times D \to B(T)$ – функция извлечения именованных сущностей из документов с помощью модели M. $L \subseteq D \times D$ – асимметричное транзитивное отношение «является откликом», определяющее на множестве документов структуру коммуникации.	T^{S+}, T^A – множества предметных и значимых общеупотребительных терминов. Предметные термины определяются внешним источником, значимые общеупотребительные – ассоциированностью с предметными и интенсивностью пользовательских реакций.

Стоит отметить, что в результате применения комбинированного метода предполагается пополнение источника знаний извлеченными терминами. Это связано с тем, что значимость, в том числе, определяется ассоциированностью, которая, в свою очередь, зависит от структуры откликов, а отклики могут содержать новые термины. Под новыми терминами в данном случае можно понимать как термины, существовавшие до этого и ставшие значимыми, так и неологизмы, появившиеся с течением времени и ставшие значимыми в рамках решаемой прикладной задачи.

2.3. Обзор ресурсов и инструментов для извлечения терминов на русском языке. Рассмотренные выше методы извлечения терминов, применительно к русскоязычным текстам, могут быть реализованы с помощью ряда ресурсов и инструментов.

2.3.1. Внешние источники знаний. В качестве внешних источников знаний могут выступать тезаурусы, онтологии, базы знаний, графы знаний и другие необходимые для решения рассматриваемой задачи ресурсы. К сожалению, подобных русскоязычных ресурсов в открытом доступе довольно мало, поэтому, зачастую, исследовательскими группами предпринимаются попытки создания их с нуля для решения той или иной задачи.

Среди существующих ресурсов общего назначения можно выделить RuWordNet [14] – наиболее распространенный русскоязычный тезаурус, представляющий собой результат трансформации тезауруса RuThes [15] в формат WordNet. По состоянию на 2016 год в нем содержатся: 29297 синсетов (наборов синонимов) существительных, 12865 синсетов прилагательных и 7636 синсетов глаголов. Всего RuWordNet содержит 111,5 тысяч слов и выражений русского языка. Также стоит отметить AGROVOC [16] – многоязычный тезаурус Продовольственной и сельскохозяйственной организации ООН, охватывающий такие области, как продовольствие, сельское хозяйство, окружающую среду и др. Он насчитывает более 41 000 понятий, включающих свыше 994 000 терминов на 42 различных языках, в том числе и на русском. Еще один известный ресурс с поддержкой русского языка – DBpedia [17] – открытая многоязычная база знаний, охватывающая множество различных предметных областей, поскольку является результатом извлечения структурированных данных из статей Википедии. DBpedia может быть полезна при решении таких задач, как интеграция данных, распознавание именованных сущностей, выявление тем и ранжирование документов.

2.3.2. Инструменты NER. Для решения задачи NER для русского языка можно выделить следующие библиотеки, реализованные на языке Python, – SpaCy [18], Natasha [19] и DeepPavlov [20].

SpaCy использует методы машинного обучения для определения в исходном тексте интервалов (непрерывных последовательностей слов), которые представляют собой сущности. В SpaCy для извлечения именованных сущностей по умолчанию применяется обученная на новостных текстах языковая модель, которая позволяет идентифицировать в тексте такие типы сущностей, как локации, организации и персоны. Также, в зависимости от специфики задачи, имеется возможность обучения собственной модели для выявления особых типов сущностей.

Библиотека Natasha, в отличие от SpaCy, ориентирована на обработку только русского языка и, соответственно, демонстрирует более точные результаты. В Natasha используется набор предопределенных шаблонов, основанных на правилах, для сопоставления и извлечения упоминаний сущностей из текста. В ней, как и в SpaCy, также используется предварительно обученная модель для извлечения стандартных типов сущностей из текста.

Библиотека DeepPavlov содержит набор предварительно обученных современных NLP моделей для анализа, в том числе, русскоязычных текстов. В реализации NER для маркировки последовательностей используется архитектура BiLSTM-CRF (Bidirectional LSTM with Conditional Random Field). Как и в рассмотренных выше библиотеках, DeepPavlov позволяет распознавать стандартные типы сущностей. Также DeepPavlov позволяет настраивать или обучать свои собственные модели NER на пользовательских размеченных наборах данных, что обеспечивает гибкость при работе с конкретными типами сущностей или предметными областями.

2.3.3. Инструменты для формирования лингвистических правил. Лингвистические правила используются для задания точных шаблонов, на основе которых из текста будут извлекаться необходимые языковые единицы. Для построения таких правил могут применяться различные парсеры, например, Томита-парсер [21] или Yargy-парсер [22], а также морфологические анализаторы, например, pymorphy2 [23].

Томита-парсер от Яндекс позволяет извлекать структурированные данные из текста на естественном языке с помощью контекстно-свободных грамматик и словарей ключевых слов. Контекстно-свободная грамматика состоит из набора правил-продукций, которые определяют, как символы и слова могут

комбинироваться для формирования действительных предложений. В данном случае правила грамматики применяются без учета контекста или окружающих слов в предложении. В основе Томита-парсера лежит алгоритм GLR-парсинга (Generalized left-to-right algorithm). Парсер также позволяет описать свои грамматики и добавлять словари для нужного языка. В Yargy-парсере – аналоге Томита-парсера для Python – правила для извлечения сущностей также описываются с помощью контекстно-свободных грамматик и словарей. Данная библиотека является частью проекта Natasha, рассмотренного выше, и, соответственно, ее основой. В Yargy-парсере реализован алгоритм Эрли – алгоритм синтаксического анализа предложения по контекстно-свободной грамматике.

Библиотека pymorphy2, используемая внутри Yargy-парсера и SpaCy для нормализации слов, может выступать в качестве самостоятельного инструмента для формирования частеречевых правил, в соответствии с которыми будут извлекаться термины конкретных частей речи и в нужной форме. Библиотека использует словари и граммемы OpenCorpora [24] для приведения слов к нормальной форме, а также получения нужной формы слов и их грамматической информации (число, род, падеж, часть речи и т.д.).

2.4. Программно-алгоритмическая реализация комбинированного метода извлечения терминов. В данной работе для совместного извлечения не только предметных, но и значимых общеупотребительных терминов использовалась комбинация последовательно применяемых методов на основе словарей, инструментов NER и правил, дополненных статистической фильтрацией. Схема комбинированного метода представлена на рисунке 2.

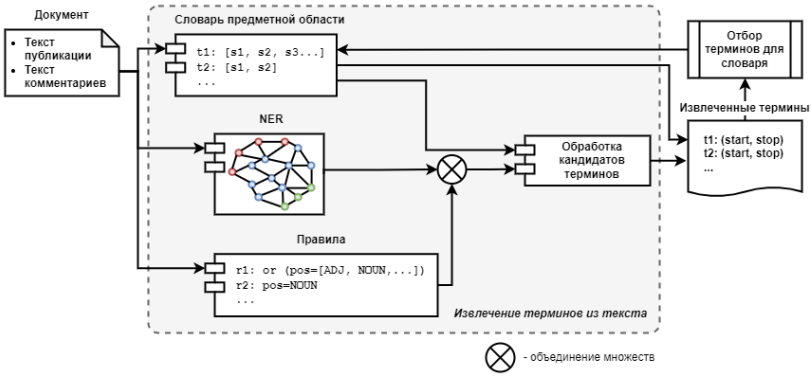

Рис. 2. Схема метода извлечения терминов

На первом этапе используется словарь для выделения предметных терминов, далее применяется NER для выделения терминов, которые могут быть как предметными (в рамках нашей задачи), так и общеупотребительными, и затем применяется набор правил для выделения общеупотребительных терминов (уни- би- и триграмм). Каждый отдельный этап («Словарь», NER, «Правила») представлен отдельным модулем в рамках разрабатываемого метода в целом. Общий алгоритм извлечения терминов описан в алгоритме 1.

Алгоритм 1. Извлечение терминов из текста предлагаемым методом

Ввод: Текстовые документы D, базовый словарь T^S, множество правил P, языковая модель M.
Вывод: Списки извлеченных терминов $T = T^{S+} \cup T^A$, где t имеет вид $(text(t), (start(t), stop(t)))$, где $text(t)$ – текстовое представление термина, полученное из документа d, $start(t)$ – позиция начала термина в документе d, $stop(t)$ – позиция окончания.
Для каждого $d_i \in D$ выполнить
Извлечь из d_i предметные термины T^{S+} с помощью словаря
Извлечь из d_i именованные сущности T^M с помощью M
Извлечь из d_i словосочетания T^{P1} с помощью Правила для выделения биграмм и триграмм (P1)
Извлечь из d_i униграммы T^{P2} с помощью Правила для выделения униграмм (P2)
Выделить из множества $T^P \cup T^M$ множество терминов T^A, ассоциированных с терминами из $T^S \cup T^{S+}$
Конец цикла

Словарь выступает инструментом однозначной идентификации терминов. Он представляет собой набор терминов предметной области, для каждого из которых хранится связанный с ним синсет, состоящий из синонимов или псевдонимов к данному термину и включающий в себя текстовое представление псевдонима и регулярное выражение для его поиска. Регулярные выражения могут быть заданы в формате точного или частичного совпадения, а также в свободной форме. Для частичного совпадения используется выделение основы текстового представления, дополняемое возможным окончанием слова в виде регулярного выражения. Начальный словарь формируется экспертом на основе анализа предметных текстов с помощью инструментов NER. В дальнейшем, при поддержке эксперта, словарь должен пополняться терминами, извлеченными в результате использования метода. Этап извлечения терминов с использованием словаря описан в алгоритме 2.

Алгоритм 2. Извлечение терминов с помощью словаря

Ввод: Документ d, базовый словарь T^S
Вывод: Список предметных терминов T^{S+}
Для каждого термина $t_i^S \in T^S$ выполнить
Если t_i^S в d
Определить границы $t_i^S.\text{start}$ и $t_i^S.\text{stop}$ в d
Записать термин t_i^S и его границы в T^{S+}
Конец цикла

Инструменты NER позволяют выделять именованные сущности, такие как имена людей, названия организаций и локаций, рассматриваемые нами как наиболее релевантные типы терминов в рамках решаемой задачи. В ходе разработки модуля был рассмотрен ряд инструментов NER: библиотека SpaCy [18], Natasha [19] и предварительно обученная модель BERT (ner_collection3_bert) [25]. Для сравнения инструментов был проведен эксперимент, в рамках которого с помощью них были извлечены именованные сущности из имеющегося датасета ВКонтакте и набора нормативно-правовых документов. Наборы извлеченных именованных сущностей были оценены экспертно, и было принято решение использовать модель BERT, т.к. она выдала наименьшее количество ложных кандидатов и ошибок. Таким образом, на вход модулю NER подается документ, где каждое его предложение анализируется моделью BERT и извлекаются все возможные термины-кандидаты и их границы в рамках текущего документа (алгоритм 3).

Алгоритм 3. Извлечение именованных сущностей с помощью языковой модели

Ввод: Документ d, языковая модель M
Вывод: Список именованных сущностей T^M
Для каждого предложения $s_i \in d$ выполнить
Извлечь из s_i именованные сущности t_j^M с помощью M
Определить границы $t_j^M.\text{start}$ и $t_j^M.\text{stop}$ в d
Записать извлеченные сущности и их границы в T^M
Конец цикла

Набор используемых правил для извлечения кандидатов терминов был создан эмпирическим путем на основе имеющихся данных и исходя из решаемой задачи. Первое правило (алгоритм 4) позволяет извлекать термины произвольной длины с минимумом предположений об их структуре (термин может состоять только из

существительных, прилагательных, причастий или числительных). Это правило основано на контекстно-свободных грамматиках, реализованных с помощью Yargy-парсера [22]. Отобранные данным правилом кандидаты фильтруются с помощью модели коллокации (реализация коллокаций из библиотеки NLTK [26]), обученной на имеющемся датасете, и дополнительно по TF для кандидата (среднее значение TF входящих в него токенов). При подсчете TF учитываются стоп-слова, список которых был получен из открытых источников и дополнен сетевым сленгом в результате анализа имеющихся данных. Максимальная длина N-грамм была ограничена до 3 токенов, что сузило круг интересующих нас словосочетаний до би- и триграмм, так как словосочетания большей длины крайне редко были полезны. Такой вывод был сделан на основе анализа результатов тестирования модуля.

Алгоритм 4. Извлечение би- и триграмм с помощью P1

Ввод: Документ d, модель коллокации C, список значений TF для документа d, Парсер
Вывод: Список словосочетаний T^{P1}
Выделить из d словосочетания N с помощью Парсера
Для каждого словосочетания $n_i \in N$ выполнить
$TF(n_i) = sum(TF(tk, n_i))/len(n_i)$, где tk – токен в n_i
Если n_i не содержит стоп-слов **и** $TF(n_i) \geq mean(TF(d))$ **и** $\mathrm{C}(n_i) \geq mean(C)$
Определить границы $n_i.\mathrm{start}$ и $n_i.\mathrm{stop}$ в d
Записать словосочетания n_i и его границы в T^{P1}
Конец цикла

Для выбора униграмм используется частеречевое правило (часть речи может быть только существительным), дополненное фильтрацией по TF с порогом в виде среднего TF по текущему документу (алгоритм 5). Минимальная длина токена была установлена на 3 символа.

Алгоритм 5. Извлечение униграмм с помощью P2

Ввод: Документ d, список значений TF для документа d
Вывод: Список униграмм T^{P2}
Для каждого токена $tk_i \in d$ выполнить
Если tk_i не является стоп-словом **и** tk_i является существительным **и** $len(tk_i) \geq 3$ **и** $TF(tk_i) \geq mean(TF(d))$
Определить границы $tk_i.\mathrm{start}$ и $tk_i.\mathrm{stop}$ в d
Записать униграмму tk_i и ее границы в T^{P2}
Конец цикла

Все извлеченные кандидаты дополнительно проходят проверку на стоп-слова: для биграмм и триграмм – это вхождение стоп-слова в кандидат, для униграмм – полное совпадение.

Набор правил может быть расширен для извлечения другой интересующей информации, например, названий улиц, номеров домов и т.д. Но на данном этапе, в рамках рассматриваемой задачи, мы ограничились описанными выше правилами для извлечения произвольных уни- би- и триграмм.

На этапе обработки терминов-кандидатов в первую очередь выполняется проверка на ассоциированность кандидата с предметным термином из словаря. Дополнительно исключаются полные совпадения извлеченных терминов на разных этапах метода, с приоритетом для вышестоящего этапа обработки. Предполагается, что, чем раньше был извлечен термин, тем больше информации о нем есть (например, термин, извлеченный словарем, однозначно идентифицирован, в отличие от униграммы, извлеченной последним этапом). Пример работы метода представлен на рисунке 8.

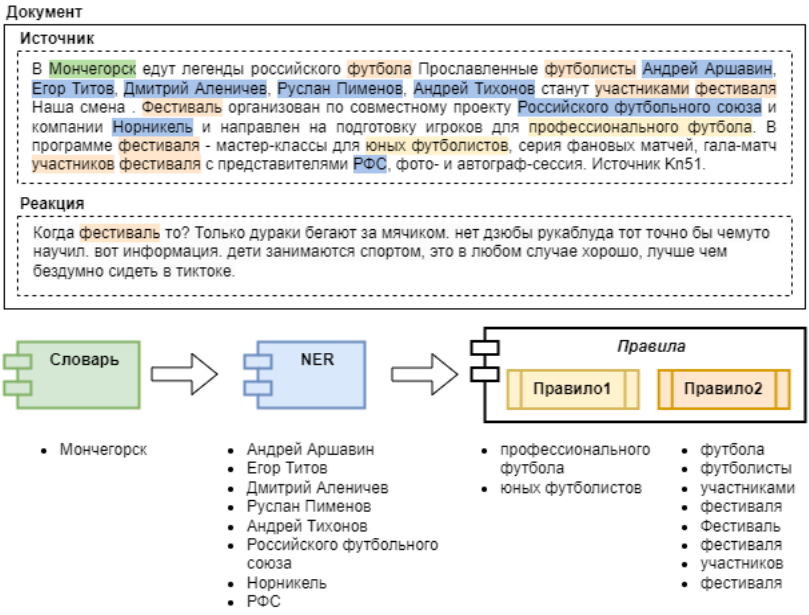

Рис. 8. Пример работы отдельных этапов метода на документе

3. Результаты. Проверка метода осуществлялась на примере прикладной задачи мониторинга социально-политической обстановки в регионе. Первоначальный словарь предметных терминов был сформирован экспертами в результате анализа набора из 18 нормативно-правовых документов, а именно уставов крупных организаций и предприятий Мурманской области.

Формирование словаря предметных терминов происходило в три этапа. Первым этапом экспертно из нормативно-правовых документов был получен исходный список терминов. Вторым этапом данный список был расширен с помощью внешнего источника знаний – Википедии. Для этого специализированным программным инструментом получения данных с веб-ресурсов были собраны связанные с исходными терминами в рамках страницы Википедии новые термины. Последним этапом экспертами, в результате анализа датасета, для каждого из терминов были сформированы связанные с ними синсеты. Таким образом, в исходный словарь было добавлено 146 предметных терминов.

Датасет, на котором проводились эксперименты, представляет собой выборку публикаций с комментариями из 8 групп социальной сети ВКонтакте по Мурманской области за приблизительно 2 года. Статистика по данному датасету приведена в таблице 2.

Таблица 2. Статистика по датасету из ВКонтакте

Показатель	Значение
Количество публикаций в датасете после очистки	50683
Количество публикаций, содержащих текст	49990
Средняя длина текста публикации (в символах)	311
Количество публикаций с комментариями	22084
Количество публикаций, содержащих текст и комментарии	21401
Количество комментариев в датасете после очистки	236444
Средняя длина одного комментария	77
Средняя длина всех комментариев к одной публикации	361

В связи с особенностями текста в датасете, необходимо было очистить его от нежелательных символов, таких как эмодзи или бессмысленные повторения различных знаков и конструкций, например, от автоматически добавляемых ссылок на пользователей в ответах на комментарии, от телефонных номеров или ссылок. На первом этапе использовались регулярные выражения для очистки от нежелательных конструкций, а на втором этапе – белый список символов, включающий в себя кириллицу, латиницу, цифры и ряд специальных символов ([^А-Яа-яЁёA-Za-z0-9,\.!?#\"\'\-—\s]). Всего

полученный датасет содержит 50683 предобработанных документа, в которых есть текст публикации и/или комментариев.

На данном этапе исследования ассоциированными считались все термины-кандидаты, выделенные в документе, в котором есть хотя бы один предметный термин, при этом интенсивность пользовательских реакций на документ-источник не учитывалась. Иначе говоря, контекст совместного употребления был задан объединением публикации с ее комментариями без учета их ветвления. Таким образом, для учета структуры коммуникации имеющийся датасет был разбит на документы, содержащие публикации и комментарии к ним.

Для оценки эффективности предлагаемого метода из датасета были выбраны и экспертно размечены 320 документов (пример на рисунке 9). Длина текста публикации и комментариев была ограничена в диапазоне от 100 до 600 символов. Такой размер документов был выбран с учетом средней длины предложения в русском языке. Исходя из решаемой прикладной задачи, эксперты размечали только те документы, в которых содержался хотя бы один термин, имеющий прямое отношение к Мурманской области. В таблице 3 представлены результаты оценки метода на размеченном датасете. В качестве оценочных метрик использовались точность, полнота и мера F1.

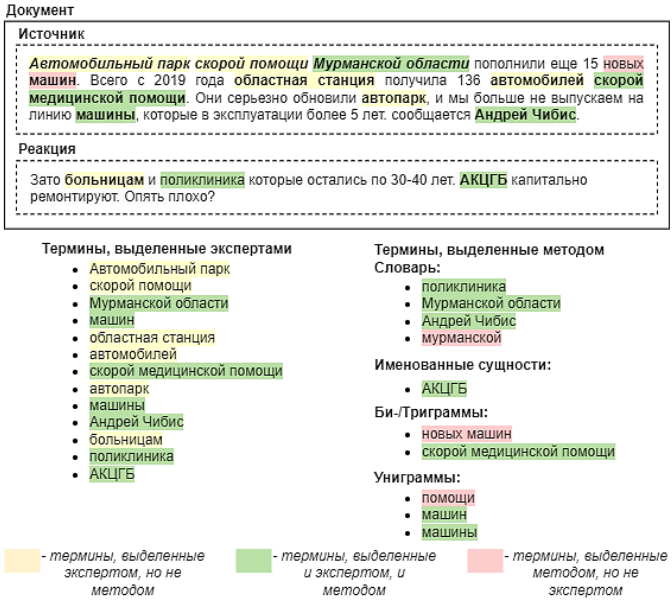

Рис. 9. Пример публикации, размеченной экспертами и методом

Таблица 3. Результаты проверки работы метода на тестовом наборе

Показатель	**Значение**
Количество терминов, размеченных экспертами	2169
Количество терминов, выделенных методом	2278
Количество терминов, выделенных и экспертами, и методом	1488
Количество терминов, выделенных экспертами, но не выделенных методом	681
Количество терминов, выделенных методом, но не выделенных экспертами	790
Precision	0,69
Recall	0,65
F1 Score	0,67

Было осуществлено сравнение эффективности разработанной реализации предложенного метода с готовыми инструментами извлечения терминов PyATE [27] и rutermextract [28] на имеющемся размеченном датасете. PyATE представляет собой Python-реализацию ряда алгоритмов извлечения терминов, среди которых наиболее точным на тестовом наборе показал себя ComboBasic [29]. Rutermextract – Python-библиотека для извлечения ключевых слов из текстов на русском языке. Результатом ее применения является упорядоченный по частоте употребления список ключевых слов. Результаты сравнения представлены в таблице 3.

Таблица 4. Результаты сравнения метода с инструментами PyATE и rutermextract

Реализация	**Precision**	**Recall**	**F1 Score**
PyATE	0,16	0,08	0,11
rutermextract	0,38	0,14	0,2
Предложенный метод	0,69	0,65	0,67

Точность предложенного метода на тестовом наборе составила 69%, что значительно превышает результативность рассмотренных в качестве аналогов готовых инструментов извлечения терминов. Принимая во внимание трудноформализуемость решаемой задачи, такая точность предложенного метода, на наш взгляд, делает его применимым для практического использования.

4. Обсуждение. Следует отметить, что рассматриваемая в данной работе задача извлечения терминов при условии множественности и динамичности рассматриваемых предметных областей является трудноформализуемой, поскольку отсутствуют объективные критерии отнесения некоторой лексической конструкции

к числу терминов (предметных или общеупотребительных). В частности, существующие публикации по данной тематике демонстрируют расхождения во мнениях экспертов, поскольку на вопрос «что является предметным термином?» не всегда имеется однозначный ответ. Возможно, единственным объективным критерием «термичности» лексической конструкции является относительная частота ее употребления в предметных текстах и текстах общей тематики [30]. Однако такой подход к идентификации терминов неприменим для рассматриваемой в данной работе категории прикладных задач. Таким образом, в нашем случае невозможно обойтись без явных или опосредованных экспертных оценок. С другой стороны, прикладной мониторинг социальных медиа предполагает обработку больших объемов данных за ограниченное время, что повышает требования к уровню автоматизации решения соответствующих задач обработки информации.

Решение задач на основе анализа данных социальных медиа встречает схожие проблемы с соотношением предметных и общеупотребительных терминов. Например, в [31] предлагается система CiCui для извлечения терминов на основе машинного обучения с целью пополнения онтологий в рамках задач управления чрезвычайными ситуациями, в которые вовлечено большое количество людей и экспертов различных предметных областей с общими целями. В качестве источника текстов для анализа авторы используют социальные медиа. Сначала предлагаемая авторами система извлекает термины-кандидаты из предварительно обработанного текста при помощи разработанных ими лингвистических шаблонов и затем очищает полученный набор кандидатов с помощью статистических классификаторов, обученных на ранее размеченных данных. В работе [32] авторами предлагается модель для извлечения терминов и классификации текстов в рамках интеллектуального анализа социальных медиа для различных задач маркетинга, анализа конкурентной информации или рынка. Применение предлагаемой авторами модели рассматривается на примере задачи выявления и классификации компонентов автомобилей на основе форумных обсуждений, хотя она может быть использована и для других типов социальных медиа, предметных областей и задач. В таком контексте задачи структура форумов, в некотором роде, дает людям небольшую долю экспертного знания, направляя их в соответствии с их нуждами на специализированный раздел форума, но в дальнейшем люди общаются, используя ту терминологию, которой владеют, и необязательно экспертную. Сначала форумные обсуждения

классифицировались на основе структуры форумов, далее из обсуждений извлекались термины-кандидаты в виде униграмм. Затем для каждой выделенной на основе структуры форумов категории выбирались наиболее значимые термины с помощью таких методов, как критерий прироста информации, критерий хи-квадрат, корреляция релевантности документа и значение выбора Робертсона. Отобранные значимые термины, категории и сами обсуждения сохраняются в базу данных обсуждений, которая используется для обучения набора бинарных классификаторов, которые, в свою очередь, используются для классификации новых публикаций.

В сфере медицины также встречаются задачи, решаемые с помощью анализа данных социальных медиа. Например, [33] предлагают три метода извлечения концептов побочных реакций на лекарства с помощью LLDA (Labeled LDA), байесовского классификатора и условных случайных полей. В этой задаче важно связать интерпретацию побочных эффектов, выраженную общеупотребительными терминами, с медицинским термином, для чего и используются предлагаемые авторами методы. В [34] авторы предлагают алгоритм MSMTC для классификации текстов социальных медиа, использующий общеупотребительные термины, относящиеся к тематике здоровья. На первом этапе тексты предобрабатываются и размечаются с помощью имеющегося словаря медицинской терминологии. Далее обе задачи, классификации текстов и извлечение терминов, решаются одновременно с использованием двухканальной структуры и обучения генеративно-состязательной сети. Первый канал обрабатывает предложения, как они есть, а второй – со скрытой медицинской терминологией, в результате получая векторные представления предложений. Результат работы второго канала используется для извлечения важных слов, которые, как предполагается, считаются общеупотребительными терминами тематики здоровья, и эти извлеченные термины добавляются в начальный словарь. Результаты работы обоих каналов используются для классификации текстов с учетом извлеченной общеупотребительной терминологии.

Предложенный в данной работе метод извлечения терминов, основанный на комбинировании различных подходов, показал достаточно высокую эффективность на рассмотренной тестовой задаче. При этом стоит заметить, что, с одной стороны, совмещение различных по своей природе методов может обеспечить синергетический эффект, но, с другой стороны, возрастают затраты на реализацию, несмотря на то, что сильные стороны одних методов

могут компенсировать слабые стороны других. Так, необходимость в постоянном пополнении словаря новыми терминами может быть компенсирована отбором кандидатов с помощью NER или специальных правил. Однако правила для отбора кандидатов придется разрабатывать на основе анализа большого количества данных и исходя из решаемой задачи, следовательно, они не всегда могут быть использованы для решения других задач этим же методом. То есть, полностью не исключается необходимость использования экспертных знаний. Также стоит отметить, что часто методы анализа текстов, используемые в качестве компонентов составной технологии, ограничены рамками языка, для которого существуют необходимые инструменты (частеречевая разметка, средства реализации правил, NER решения) и ресурсы (размеченные корпуса, тезаурусы и т.п.).

Метод применим как для социальных медиа, так и для других источников, где преобладают текстовые данные и прослеживается структура «документ-отклик». При этом наличие в рамках социальной медиа или другого источника индикаторов пользовательской реакции не является принципиальным, но потенциально снижает качество извлечения терминов, поскольку значение обобщенной оценки интенсивности реакции на документ для всех документов будет идентичным (нулевым).

В качестве перспективы развития предложенного метода можно рассматривать оперирование не только «плоскими», но и вложенными сущностями. В частности, на этапе обработки кандидатов существует возможность решения задачи выделения вложенных терминов или сущностей (nested NER). Пример такой возможности можно наблюдать на рисунке 9, где можно было дополнительно выделить, в том числе, термин «Автомобильный парк скорой помощи Мурманской области».

5. Заключение. В данной работе рассмотрена проблема автоматизированного извлечения терминов из естественно-языковых текстов для последующего их использования при построении формализованных систем знаний (онтологий, тезаурусов, графов знаний) в рамках задачи мониторинга тематических обсуждений в социальных медиа.

Извлечение терминов является ключевой проблемой автоматизированного построения систем знаний на основе текстов на естественном языке, поскольку обеспечивает формирование базовой системы понятий, используемой затем в прикладных задачах интеллектуальной обработки информации. В то же время, данная задача относится к категории трудноформализуемых, поскольку

критерий качества ее решения зависит от прикладной задачи, в которой предполагается использовать результат.

Рассматриваемая в данной работе прикладная задача характеризуется тем, что в формируемую систему знаний необходимо включать как понятия из нескольких различных предметных областей, так и значимые общеупотребительные понятия, используемые аудиторией социальных медиа в рамках тематических обсуждений. Кроме того, формируемая система знаний является динамичной как с точки зрения состава охватываемых ею предметных областей, так и состава релевантных понятий, подлежащих включению в систему.

Для автоматизированного извлечения терминов из текстов в настоящее время создано большое количество методов и технологий, которые можно разделить на три основные категории – методы на основе правил, на основе словарей или внешних источников знаний, а также статистические. Однако прямое использование упомянутых методов для решения рассматриваемой здесь категории задач затруднительно, поскольку, как правило, все они предполагают наличие некоторой единственной предметной области, в рамках которой формируется множество терминов, обозначающих соответствующие понятия. Помимо этого, такие методы не принимают во внимание особенности, присущие текстам социальных медиа. Таким образом, применение только классических методов не покрывает все интересующие нас термины и, следовательно, не обеспечивает достаточной полноты в рамках решаемой задачи.

Для решения задачи извлечения терминов в указанных выше условиях, характерных для рассматриваемой категории прикладных задач, в данной работе предложен комбинированный метод, совмещающий в себе подходы на основе внешних источников знаний, инструментов NER и правил. Особенностью метода является учет структуры коммуникации и пользовательских реакций при принятии решения о включении некоторого термина в число значимых. Результаты экспериментов на тестовом наборе подтверждают эффективность предложенной комбинации подходов к извлечению терминов, позволяющей извлечь как предметные, так и значимые общеупотребительные термины, важные для задачи мониторинга и анализа тематических обсуждений в социальных медиа. Разработанный метод значительно превосходит по точности существующие инструменты извлечения терминов, реализованные в виде программных библиотек на языке Python.

Литература

1. Матусевич А.А. Общение в социальных сетях: прагматический, коммуникативный, лингвостилистический аспекты характеристики: дис. ... канд. филол. наук: 10.02.01. Киров. 2016. 190 с.
2. Mykowiecka A., Marciniak M., Kupsc A. Rule-based information extraction from patients' clinical data // Journal of Biomedical Informatics. 2009. vol. 42. no 5. pp. 923–936.
3. Golik W., Bossy R., Ratkovic Z., Nedellec C. Improving term extraction with linguistic analysis in the biomedical domain // RCS. 2013. vol. 70. no. 1. pp. 157–172.
4. Zhou X., Zhang X., Hu X. MaxMatcher: Biological Concept Extraction Using Approximate Dictionary Lookup // PRICAI 2006: Trends in Artificial Intelligence Lecture Notes in Computer Science. 2006. pp. 1145–1149.
5. Yehia E., Boshnak H., AbdelGaber S., Abdo A., Elzanfaly. D.S. Ontology-based clinical information extraction from physician's free-text notes // Journal of Biomedical Informatics. 2019. vol. 98. no. 103276.
6. Lomov P., Malozemova M., Shishaev M. Training and application of neural-network language model for ontology population // Software engineering perspectives in intelligent systems: Proceedings of 4th Computational Methods in Systems and Software. 2020. vol. 1295. pp. 919–926.
7. Пимешков В.К., Диковицкий В.В., Шишаев М.Г. Формирование тренировочных наборов данных для нейросетевого классификатора в задаче извлечения понятий и отношений из естественно-языковых текстов // Сборник Региональной научно-практической конференции-студенческой научной школы филиала МАГУ в г. Апатиты. 2021. С. 158–170.
8. Pazienza M.T., Pennacchiotti M., Zanzotto F.M. Terminology Extraction: An Analysis of Linguistic and Statistical Approaches // Knowledge Mining. Studies in Fuzziness and Soft Computing. 2005. pp. 255–279.
9. Heid U. Extracting terminologically relevant collocations from German technical texts // Terminology and Knowledge Engineering Proceedings. 1999. vol. 99. pp. 242–255.
10. Zhang S., Elhadad N. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts // Journal of Biomedical Informatics. 2013. vol. 46. no 6. pp. 1088–1098.
11. Lopez P., Romary L. HUMB: Automatic Key Term Extraction from Scientific Articles in GROBID // Proceedings of the 5th International Workshop on Semantic Evaluation. 2010. pp. 248–251.
12. GROBID. URL: https://github.com/kermitt2/grobid (дата обращения: 13.11.2023).
13. Lopez P., Romary L. GRISP: A Massive Multilingual Terminological Database for Scientific and Technical Domains // Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). 2010. pp. 2269–2276.
14. RuWordNet. URL: https://ruwordnet.ru/ru/ (дата обращения: 13.11.2023).
15. RuThes. URL: http://www.labinform.ru/pub/ruthes/ (дата обращения: 13.11.2023).
16. AGROVOC. URL: https://www.fao.org/agrovoc/ (дата обращения: 13.11.2023).
17. DBpedia. URL: https://www.dbpedia.org/ (дата обращения: 13.11.2023).
18. SpaCy. URL: https://spacy.io/ (дата обращения: 07.12.2023).
19. Natasha. URL: https://github.com/natasha/natasha (дата обращения: 24.10.2023).
20. DeepPavlov. URL: https://docs.deeppavlov.ai/en/master/ (дата обращения: 07.12.2023).
21. Томита-парсер. URL: https://yandex.ru/dev/tomita/ (дата обращения: 13.11.2023).
22. Yargy parser. URL: https://github.com/natasha/yargy (дата обращения: 24.10.2023).
23. Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts. Communications in Computer and Information Science. 2015. pp. 320–332.
24. OpenCorpora. URL: https://opencorpora.org/ (дата обращения: 13.11.2023).

25. BERT NER-models. URL: https://docs.deeppavlov.ai/en/master/features/models/NER.html#6.-Models-list (дата обращения: 24.10.2023).
26. Bird S., Klein E., Loper E. Natural Language Processing with Python // O'Reilly Media Inc. 2009. 482 p.
27. Lu K. kevinlu1248/pyate: Python Automated Term Extraction (Version v0.5.3). Zenodo. 2021. DOI: 10.5281/zenodo.5039289.
28. Rutermextract. URL: https://pypi.org/project/rutermextract/ (дата обращения: 07.12.2023).
29. Astrakhantsev N. ATR4S: toolkit with state-of-the-art automatic terms recognition methods in Scala // Language Resources and Evaluation. 2018. vol. 52. no. 3. pp. 853–872.
30. Hatty A., Schlechtweg D., Dorna M., im Walde S.S. Predicting Degrees of Technicality in Automatic Terminology Extraction // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. pp. 2883–2889.
31. Zhang X., Panizzon R., Musacchio M., Ahmad K. Terminology Extraction for and from Communications in Multi-disciplinary Domains // Proceedings of the LREC 2016 Workshop "EMOT: Emotions, Metaphors, Ontology and Terminology during Disasters". 2016. pp. 34–39.
32. Abrahams A.S., Jiao J., Fan W., Wang G.A., Zhang Z. What's buzzing in the blizzard of buzz? Automotive component isolation in social media postings // Decision Support Systems. 2013. vol. 55. no. 4. pp. 871–882.
33. Yates A., Goharian N., Frieder O. Extracting Adverse Drug Reactions from Social Media // Proceedings of the AAAI Conference on Artificial Intelligence. 2015. vol. 29. no. 1. DOI: 10.1609/aaai.v29i1.9527.
34. Liu K., Chen L. Medical Social Media Text Classification Integrating Consumer Health Terminology // IEEE Access. 2019. vol. 7. pp. 78185–78193.

Пимешков Вадим Константинович — аспирант, стажер-исследователь, лаборатория информационных технологий управления региональным развитием, ИИММ КНЦ РАН. Область научных интересов: обработка естественного языка, извлечение знаний из естественно языковых текстов. Число научных публикаций — 7. pimeshkov@iimm.ru; улица Ферсмана, 24А, 184209, Апатиты, Россия; р.т.: +7(815)557-9689.

Никонорова Марина Леонидовна — аспирант, инженер-исследователь, лаборатория информационных технологий управления региональным развитием, ИИММ КНЦ РАН. Область научных интересов: анализ естественного языка, извлечение знаний, онтологии. Число научных публикаций — 11. nikonorova@iimm.ru; улица Ферсмана, 24А, 184209, Апатиты, Россия; р.т.: +7(815)557-9689.

Шишаев Максим Геннадьевич — д-р техн. наук, профессор, главный научный сотрудник, руководитель лаборатории, лаборатория информационных технологий управления региональным развитием, ИИММ КНЦ РАН; профессор кафедры, кафедра информатики и вычислительной техники, филиал МАУ в г. Апатиты. Область научных интересов: информационные системы, региональное развитие, инженерия знаний, искусственный интеллект, машинное обучение, интеллектуальный анализ текстов. Число научных публикаций — 167. shishaev@iimm.ru; улица Ферсмана, 24А, 184209, Апатиты, Россия; р.т.: +7(815)557-9248.

Поддержка исследований. Работа выполнена при поддержке Министерства науки и высшего образования РФ (№122022800551-0, FMEZ-2022-0007). Авторы выражают благодарность Вишнякову Ивану Геннадьевичу, системному администратору ИИММ КНЦ РАН, за предоставленный экспериментальный датасет.

DOI 10.15622/ia.23.4.7

V. PIMESHKOV, M. NIKONOROVA, M. SHISHAEV

A COMBINED TERM EXTRACTION METHOD FOR THE PROBLEM OF MONITORING THEMATIC DISCUSSIONS IN SOCIAL MEDIA

Pimeshkov V., Nikonorova M., Shishaev M. **A Combined Term Extraction Method for the Problem of Monitoring Thematic Discussions in Social Media.**

Abstract. Term extraction is an important stage in the automated construction of knowledge systems based on natural language texts, since it provides the formation of a basic concept system, which is then used in applied problems of intellectual information processing. The article discusses the problem of automated extraction of terms from natural language texts for their further use in the construction of formalized knowledge systems (ontologies, thesauruses, knowledge graphs) within the problem of monitoring thematic discussions in social media. This problem is characterized by the need to include in the formed knowledge system both concepts from several different domains, and some general concepts used by the audience of social media within thematic discussions. In addition, the generated knowledge system is dynamic both in terms of the composition of the domains it covers and the composition of relevant concepts to be included in the system. The use of existing classical methods for term extraction in this case is difficult, since they are focused on extracting terms within one domain. Based on this, to solve the problem under consideration, a combined method is proposed, combining approaches based on dictionaries, NER tools and rules. The results of the experiments demonstrate the effectiveness of the proposed combination of approaches to term extraction, which makes it possible to extract terms for the problem of monitoring and analyzing thematic discussions in social media. The developed method significantly exceeds the precision of the considered term extraction tools. As a further direction of research, the possibility of developing a method for solving the problem of identifying nested terms or entities is considered.

Keywords: text mining, term extraction, social media, knowledge extraction.

References

1. Matusevich A.A. Obshhenie v social'nyh setjah: pragmaticheskij, kommunikativnyj, lingvostilisticheskij aspekty harakteristiki [Communication in social networks: pragmatic, communicative, linguostylistic aspects of characteristics]. Kirov. 2016. 190 p. (In Russ.).
2. Mykowiecka A., Marciniak M., Kupsc A. Rule-based information extraction from patients' clinical data.Journal of Biomedical Informatics. 2009. vol. 42. no 5. pp. 923–936.
3. Golik W., Bossy R., Ratkovic Z., Nedellec C. Improving term extraction with linguistic analysis in the biomedical domain. RCS. 2013. vol. 70. no. 1. pp. 157–172.
4. Zhou X., Zhang X., Hu X. MaxMatcher: Biological Concept Extraction Using Approximate Dictionary Lookup. PRICAI 2006: Trends in Artificial Intelligence Lecture Notes in Computer Science. 2006. pp. 1145–1149.
5. Yehia E., Boshnak H., AbdelGaber S., Abdo A., Elzanfaly. D.S. Ontology-based clinical information extraction from physician's free-text notes. Journal of Biomedical Informatics. 2019. vol. 98. no. 103276.
6. Lomov P., Malozemova M., Shishaev M. Training and application of neural-network language model for ontology population. Software engineering perspectives

in intelligent systems: Proceedings of 4th Computational Methods in Systems and Software. 2020. vol. 1295. pp. 919–926.
7. Pimeshkov V.K. Dikovitsky V.V. Shishaev M.G. [Formation of training data sets for a neural network classifier in the problem of extracting concepts and relations from natural language texts] Sbornik Regional'noj nauchno-prakticheskoj konferencii-studencheskoj nauchnoj shkoly filiala MAGU v g. Apatity [Collection of the Regional scientific and practical conference-student scientific school of Apatity branch of MASU]. 2021. pp. 158-170. (In Russ.).
8. Pazienza M.T., Pennacchiotti M., Zanzotto F.M. Terminology Extraction: An Analysis of Linguistic and Statistical Approaches. Knowledge Mining. Studies in Fuzziness and Soft Computing. 2005. pp. 255–279.
9. Heid U. Extracting terminologically relevant collocations from German technical texts. Terminology and Knowledge Engineering Proceedings. 1999. vol. 99. pp. 242–255.
10. Zhang S., Elhadad N. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. Journal of Biomedical Informatics. 2013. vol. 46. no 6. pp. 1088–1098.
11. Lopez P., Romary L. HUMB: Automatic Key Term Extraction from Scientific Articles in GROBID. Proceedings of the 5th International Workshop on Semantic Evaluation. 2010. pp. 248–251.
12. GROBID. Available at: https://github.com/kermitt2/grobid (accessed 13.11.2023).
13. Lopez P., Romary L. GRISP: A Massive Multilingual Terminological Database for Scientific and Technical Domains. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). 2010. pp. 2269–2276.
14. RuWordNet. Available at: https://ruwordnet.ru/ru/ (accessed 13.11.2023).
15. RuThes. Available at: http://www.labinform.ru/pub/ruthes/ (accessed 13.11.2023).
16. AGROVOC. Available at: https://www.fao.org/agrovoc/ (accessed 13.11.2023).
17. DBpedia. Available at: https://www.dbpedia.org/ (accessed 13.11.2023).
18. SpaCy. Available at: https://spacy.io/ (accessed 07.12.2023).
19. Natasha. Available at: https://github.com/natasha/natasha (accessed 24.10.2023).
20. DeepPavlov. Available at: https://docs.deeppavlov.ai/en/master/ (accessed 07.12.2023).
21. Томита-парсер. Available at: https://yandex.ru/dev/tomita/ (accessed 13.11.2023).
22. Yargy parser. Available at: https://github.com/natasha/yargy (accessed 24.10.2023).
23. Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages. Analysis of Images, Social Networks and Texts. Communications in Computer and Information Science. 2015. pp. 320–332.
24. OpenCorpora. Available at: https://opencorpora.org/ (accessed 13.11.2023).
25. BERT NER-models. Available at: https://docs.deeppavlov.ai/en/master/features/models/NER.html#6.-Models-list (accessed 24.10.2023).
26. Bird S., Klein E., Loper E. Natural Language Processing with Python. O'Reilly Media Inc. 2009. 482 p.
27. Lu K. kevinlu1248/pyate: Python Automated Term Extraction (Version v0.5.3). Zenodo. 2021. DOI: 10.5281/zenodo.5039289.
28. Rutermextract. Available at: https://pypi.org/project/rutermextract/ (accessed 07.12.2023).
29. Astrakhantsev N. ATR4S: toolkit with state-of-the-art automatic terms recognition methods in Scala. Language Resources and Evaluation. 2018. vol. 52. no. 3. pp. 853–872.

30. Hatty A., Schlechtweg D., Dorna M., im Walde S.S. Predicting Degrees of Technicality in Automatic Terminology Extraction. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. pp. 2883–2889.
31. Zhang X., Panizzon R., Musacchio M., Ahmad K. Terminology Extraction for and from Communications in Multi-disciplinary Domains. Proceedings of the LREC 2016 Workshop "EMOT: Emotions, Metaphors, Ontology and Terminology during Disasters". 2016. pp. 34–39.
32. Abrahams A.S., Jiao J., Fan W., Wang G.A., Zhang Z. What's buzzing in the blizzard of buzz? Automotive component isolation in social media postings. Decision Support Systems. 2013. vol. 55. no. 4. pp. 871–882.
33. Yates A., Goharian N., Frieder O. Extracting Adverse Drug Reactions from Social Media. Proceedings of the AAAI Conference on Artificial Intelligence. 2015. vol. 29. no. 1. DOI: 10.1609/aaai.v29i1.9527.
34. Liu K., Chen L. Medical Social Media Text Classification Integrating Consumer Health Terminology. IEEE Access. 2019. vol. 7. pp. 78185–78193.

Pimeshkov Vadim — Postgraduate student, research intern, Laboratory of information technologies for regional development management, IIMM KSC RAS. Research interests: natural language processing, knowledge extraction from natural language texts. The number of publications — 7. pimeshkov@iimm.ru; 24A, Fersman St., 184209, Apatity, Russia; office phone: +7(815)557-9689.

Nikonorova Marina — Postgraduate student, research engineer, Laboratory of information technologies for regional development management, IIMM KSC RAS. Research interests: natural language analysis, knowledge extraction, ontologies. The number of publications — 11. nikonorova@iimm.ru; 24A, Fersman St., 184209, Apatity, Russia; office phone: +7(815)557-9689.

Shishaev Maxim — Ph.D., Dr.Sci., Professor, Chief researcher, Laboratory of information technologies for regional development management, IIMM KSC RAS; Professor of the department, Informatics and computing engineering department, Apatity branch of MAU. Research interests: information systems, regional development, knowledge engineering, artificial intelligence, machine learning, natural language processing. The number of publications — 167. shishaev@iimm.ru; 24A, Fersman St., 184209, Apatity, Russia; office phone: +7(815)557-9248.

Acknowledgements. This work was supported by the Ministry of Science and Higher Education of the Russian Federation (No.122022800551-0, FMEZ-2022-0007). The authors express their gratitude to Ivan Vishnyakov, System administrator of IIMM KSC RAS, for providing the experimental dataset.