КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2021 Т. 13 № 6 С. 1291–1315
DOI: 10.20537/2076-7633-2021-13-6-1291-1315

Ки&М

МОДЕЛИ ЭКОНОМИЧЕСКИХ И СОЦИАЛЬНЫХ СИСТЕМ

УДК: 519.254

Обзор современных технологий извлечения знаний из текстовых сообщений

А. А. Мусаев[1,a], Д. А. Григорьев[2,b]

[1]Санкт-Петербургский государственный технологический институт, Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН),
Россия, 199178, г. Санкт-Петербург, ВО, 14 линия, д. 39
[2]Санкт-Петербургский государственный университет (СПБГУ),
Россия, 199034, г. Санкт-Петербург, Университетская набережная, 7–9

E-mail: [a] amusaev@technolog.edu.ru, [b] d.a.grigoriev@spbu.ru

Получено 20.04.2021, после доработки — 24.10.2021.
Принято к публикации 26.10.2021.

Решение общей проблемы информационного взрыва связано с системами автоматической обработки цифровых данных, включая их распознавание, сортировку, содержательную обработку и представление в виде, приемлемом для восприятия человеком. Естественным решением является создание интеллектуальных систем извлечения знаний из неструктурированной информации. При этом явные успехи в области обработки структурированных данных контрастируют со скромными достижениями в области анализа неструктурированной информации, в частности в задачах обработки текстовых документов. В настоящее время данное направление находится в стадии интенсивных исследований и разработок. Данная работа представляет собой системный обзор международных и отечественных публикаций, посвященных ведущему тренду в области автоматической обработки потоков текстовой информации, а именно интеллектуальному анализу текстов или Text Mining (TM). Рассмотрены основные задачи и понятия TM, его место в области проблемы искусственного интеллекта, а также указаны сложности при обработке текстов на естественном языке (NLP), обусловленные слабой структурированностью и неоднозначностью лингвистической информации. Описаны стадии предварительной обработки текстов, их очистка и селекция признаков, которые, наряду с результатами морфологического, синтаксического и семантического анализа, являются компонентами TM. Процесс интеллектуального анализа текстов представлен как отображение множества текстовых документов в «знания», т. е. в очищенную от избыточности и шума совокупность сведений, необходимых для решения конкретной прикладной задачи. На примере задачи трейдинга продемонстрирована формализация принятия торгового решения, основанная на совокупности аналитических рекомендаций. Типичными примерами TM являются задачи и технологии информационного поиска (IR), суммаризации текста, анализа тональности, классификации и кластеризации документов и т. п. Общим вопросом для всех методов TM является выбор типа словоформ и их производных, используемых для распознавания контента в последовательностях символов NL. На примере IR рассмотрены типовые алгоритмы поиска, основанные на простых словоформах, фразах, шаблонах и концептах, а также более сложные технологии, связанные с дополнением шаблонов синтаксической и семантической информацией. В общем виде дано описание механизмов NLP: морфологический, синтаксический, семантический и прагматический анализ. Приведен сравнительный анализ современных инструментов TM, позволяющий осуществить выбор платформы, исходя из особенности решаемой задачи и практических навыков пользователя.

Ключевые слова: извлечение знаний, извлечение информации, обработка естественного языка, машинное обучение, семантическое аннотирование

Работа выполнена при частичной финансовой поддержке грантов РФФИ (№№ 19-08-00989, 20-08-01046), в рамках бюджетной темы № 0073-2019-0004 (А. А. Мусаев), а также при финансовой поддержке гранта Санкт-Петербургского государственного университета № 60419633 и в рамках Программы исследований по эконометрике и бизнес-аналитике Центра ЦЭБА СПбГУ (Д. А. Григорьев).

COMPUTER RESEARCH AND MODELING
2021 VOL. 13 NO. 6 P. 1291–1315
DOI: 10.20537/2076-7633-2021-13-6-1291-1315

КиМ

MODELS OF ECONOMIC AND SOCIAL SYSTEMS

UDC: 519.254

Extracting knowledge from text messages: overview and state-of-the-art

A. A. Musaev[1,a], D. A. Grigoriev[2,b]

[1]St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS),
39 Linia 14-th, VO, St. Petersburg, 199178, Russia
[2]Saint-Petersburg State University (SPBU),
7/9 Universitetskaya Emb., St Petersburg 199034, Russia

E-mail: [a] amusaev@technolog.edu.ru, [b] d.a.grigoriev@spbu.ru

Received 20.04.2021, after completion — 24.10.2021.
Accepted for publication 26.10.2021.

In general, solving the information explosion problem can be delegated to systems for automatic processing of digital data. These systems are intended for recognizing, sorting, meaningfully processing and presenting data in formats readable and interpretable by humans. The creation of intelligent knowledge extraction systems that handle unstructured data would be a natural solution in this area. At the same time, the evident progress in these tasks for structured data contrasts with the limited success of unstructured data processing, and, in particular, document processing. Currently, this research area is undergoing active development and investigation. The present paper is a systematic survey on both Russian and international publications that are dedicated to the leading trend in automatic text data processing: Text Mining (TM). We cover the main tasks and notions of TM, as well as its place in the current AI landscape. Furthermore, we analyze the complications that arise during the processing of texts written in natural language (NLP) which are weakly structured and often provide ambiguous linguistic information. We describe the stages of text data preparation, cleaning, and selecting features which, alongside the data obtained via morphological, syntactic, and semantic analysis, constitute the input for the TM process. This process can be represented as mapping a set of text documents to «knowledge». Using the case of stock trading, we demonstrate the formalization of the problem of making a trade decision based on a set of analytical recommendations. Examples of such mappings are methods of Information Retrieval (IR), text summarization, sentiment analysis, document classification and clustering, etc. The common point of all tasks and techniques of TM is the selection of word forms and their derivatives used to recognize content in NL symbol sequences. Considering IR as an example, we examine classic types of search, such as searching for word forms, phrases, patterns and concepts. Additionally, we consider the augmentation of patterns with syntactic and semantic information. Next, we provide a general description of all NLP instruments: morphological, syntactic, semantic and pragmatic analysis. Finally, we end the paper with a comparative analysis of modern TM tools which can be helpful for selecting a suitable TM platform based on the user's needs and skills.

Keywords: text mining, information extraction, natural language processing, machine learning, semantic annotations

Citation: *Computer Research and Modeling*, 2021, vol. 13, no. 6, pp. 1291–1315 (Russian).

The work is partially supported by the Russian Foundation for Basic Research (grants 19-08-00989, 20-08-01046), state research 0073-2019-0004 (A. A. Musaev) and by the SPBU grant (project No. 60419633) as well as within the framework of the CEBA Center Research Program at SPBU (D. A. Grigoriev).

1. Введение

Очередным нерешенным вызовом XXI века стала проблема информационного взрыва, обусловленная экспоненциальным ростом объема генерируемой человечеством информации. Объем цифровых данных возрастает ежегодно на 30 %. Особенно высокая скорость нарастания объема информации имеет место во Всемирной сети и на электронных носителях [Копчёнова, Орлова, Селезнёва, 2013; Kitsuregawa, Nishida, 2010].

Статистические исследования показали, что объем цифровой информации приближенно следует закону Мура и удваивается каждые 18 месяцев. При этом до 95 % этого потока состоит из неструктурированных данных [Gantz, Reinsel, 2011]. Расширяется сфера понимания того, что информация является главным ресурсом человечества, и если можно надеяться на разрешение мальтузианского демографического кризиса за счет снижения скорости прироста рождаемости с ростом индивидуального материального благосостояния, как это произошло в высокоразвитых странах, то в отношении информационного кризиса вопрос остается открытым.

Естественным направлением для решения данной проблемы является создание интеллектуальных систем извлечения знаний из неструктурированной информации. В частности, необходимость автоматического извлечения полезной информации (или знаний) из потока текстовых документов привела к возникновению научного направления, получившего наименование *Text Mining* (TM), или раскопок знаний в текстовой информации. Обычно TM рассматривают как составную часть более общей концепции интеллектуального анализа данных (или *Data Mining*). С другой стороны, очевидна связь TM с другим современным трендом в области *информационных технологий* (ИТ) — *проблемой больших данных* (Big Data).

Задача *извлечения знаний* (Knowledge Extraction, KE) из потоков текстовой информации оказалась настолько актуальной, что породила множество отечественных и зарубежных публикаций по данному вопросу. В связи с этим возникла потребность в критической систематизации материалов, посвященных исследованию современных технологий автоматического извлечения знаний из текстовых сообщений. Системность изложения в данной статье следует подходу [Барсегян и др., 2008], где описаны постановка задачи KE и различные методы ее решения. В то же время несомненные успехи в области обработки текстовой информации, представленной *на естественном языке* (NLP, natural language processing), делают актуальным анализ методов решения задач TM на основе NLP. В работах [Фомичева, Магомедов, Викулина, 2019; Тимофеев, Лебединская, 2017] подчеркивается важность применения анализа текстовых данных при контроле и выработке управленческих решений, в то время как авторы [Гордиенко, Паненко, 2018] адаптируют применение технологий и анализа больших данных к задачам экономической деятельности. Данная статья содержит описание универсальных методов TM и примеры извлечения классификационных признаков из финансовых текстов. Наконец, существующие учебные пособия [Наумов, 2020; Дьяконов, 2010] фокусируются на применении одного или нескольких математических и программных инструментов для задач TM, оставляя за рамками существующий спектр прикладного *программного обеспечения* (ПО). Краткий обзор программного инструментария для решения задач KE завершает настоящий обзор современных публикаций в области TM.

2. Основные определения и задачи Text Mining

Любые исследовании в области ИТ сталкиваются с проблемой неопределенности и нечеткости при определении базовых понятий: информация, данные и знания. Отечественные и зарубежные стандарты [ГОСТ 7.0-99, 2000; ISO/IEC, 2015] определяют информацию через тавтологические понятия типа «сведения», «представления» и т. п. Сложность этого определения привела к тому, что некоторые авторы вообще считают не нужным вводить это определение.

Так, Н. Винер дает такое определение: «*Информация — это не материя и не энергия, информация — это информация*» [Винер, 1983]. Представляется интересным рассмотреть информацию как нематериальные последовательности, формирующие упорядочивающие противотечения в общем потоке возрастания энтропии материального мира.

В контексте ТМ понятие исходной информации конкретизируется в форме текстовых документов, образующих набор неструктурированных или плохо структурированных данных. Задача ТМ состоит в извлечении знаний, под которыми будем понимать полезную информацию, необходимую для решения конкретных прикладных задач. Таким образом, понятие *знание* обусловлено содержательным аспектом метазадачи, в интересах которой осуществляется обработка текстовой информации.

В качестве сквозного примера рассмотрим задачу управления финансовыми активами (трейдинга) на электронных рынках капитала на основе результатов фундаментального анализа текущей финансовой, экономической и политической ситуации. Исходными данными являются потоки электронных текстовых сообщений (аналитических обзоров, прогнозов, новостной информации и т. п.), публикуемых аналитиками на сайтах финансовых компаний. Задачей автоматического извлечения знаний в этом случае является формирование прогностических оценок динамики котировок используемых финансовых инструментов. Иными словами, знанием здесь является прогноз котировок, осуществляемый в интересах вышестоящей метасистемы управления финансовыми активами. В работе [Xing, Cambria, Welsch, 2018] приведены примеры управляющих решений подобной метасистемы на основе ежедневных прогнозов по завтрашнему значению цены, составленных на основе извлечения признаков из релевантных текстовых источников. Первый вариант (buy up/sell down) — покупать, когда предсказанная цена выше текущей, иначе — продавать. Вторая тактика short-term reversal подразумевает собой арбитраж при временной раскорреляции поведения цены и новостного фона, например, из-за выхода на рынок крупного игрока.

В зависимости от вида извлекаемых и формируемых знаний обычно различают следующие основные задачи ТМ:

- собственно извлечение знаний, т. е. выявление в текстах сведений, необходимых для решения тех или иных задач;
- поиск нужной информации в тексте (retrieval information);
- классификация и сортировка документов в соответствии с априорно заданными классами или одновременно с формированием множества классов;
- суммаризация (summarization) или аннотирование, т. е. создание краткой справки, отражающей содержание документа.

3. Математический инструментарий Text Mining

Различные технологии выявления знаний в текстовых сообщениях можно рассматривать в виде подразделов таких направлений теории *искусственного интеллекта* (artificial intelligence, AI), как обработка текстов, написанных на естественном языке, и интеллектуальный анализ данных. Однако справедлива и обратная схема, когда ТМ рассматривается как некоторая прикладная надстройка, использующая математический арсенал одновременно обоих указанных направлений. В свою очередь, NLP опирается на методы математической лингвистики и статистического анализа данных.

Широкое использование математических методов классификации указывает на пересечение инструментария задач TM и *машинного обучения* (Machine Learning, ML), причем в зависимости от особенностей решаемой задачи используются методы обучения как с учителем, так и без учителя, т. е. на основе кластеризации данных.

4. Text Mining: проблемы реализации

Основная проблема извлечения знаний из текстовой информации, сформированной на основе *естественных языков* (natural language, NL), состоит в слабой структурированности и неоднозначности лингвистической информации. Одни и те же слова и фразы будут обладать различной семантикой.

Другая проблема TM и NLP связана с избыточностью вербальной информации. Необходимость в такой избыточности обусловлена потребностью в помехоустойчивости передачи информации, осуществляемой с помощью акустических колебаний среды погружения. Однако с точки зрения TM избыточная текстовая информация представляет собой шумовую составляющую и приводит к обратному эффекту, т. е. к снижению вероятности правильного извлечения знаний. Примерами такой избыточности могут служить использование нескольких слов для выражения одной мысли, элементы множественного согласования в морфологии или множественные признаки для различения фонем. Величина избыточности разных языков мира колеблется в пределах 70–80 % [Гуларян, 2010].

Третья проблема TM состоит в сложности формирования обучающих *баз данных* (БД). Исходные данные необходимо собирать с множества разнородных сайтов, тексты могут различаться естественным языком, форматом (HTML, MS Word, PDF, CSS и т. п.), степенью структурированности и т. п. Иными словами, имеет место известная проблема *слияния разнородных данных* (Data Fusion).

Существенная сложность формирования обучающих данных состоит в том, что значительная часть текстов, отобранных для ML по названию или аннотации публикаций, не содержат никакой полезной информации. Так, например, финансовые аналитики, по вполне понятным причинам, пишут тексты, из которых, в силу их двусмысленности и размытости, даже человеку невозможно извлечь какие-то прогностические оценки. Здесь нет ничего нового, этими приемами пользовались и египетские жрецы, и оракулы Древней Греции.

С лингвистической точки зрения сложность автоматического извлечения знания состоит в неоднозначности NL. В частности, возникают проблемы с раскрытием анафор (неоднозначность при использовании местоимений), изменениями в порядке слов в предложении, неологизмами, восприятием синонимов и т. д. Широкое использование математических методов классификации указывает на пересечение инструментария задач TM и Data Mining, причем в зависимости от особенностей решаемой задачи используются методы обучения как с учителем, так и без учителя.

5. Технические аспекты Text Mining

Прежде чем приступить собственно к задаче извлечения знаний из текстовых сообщений, необходимо осуществить ряд технических манипуляций, включающих в себя сбор документов, приведение их к единому формату и предобработку.

Предобработка (preprocessing) предназначена для устранения избыточности и преобразования документа к виду, приемлемому для дальнейшего семантического анализа. Предобработка включает в себя:

- *токенизацию*, т. е. представление документа в форме унифицированных строковых последовательностей;

- *стемминг* (или сходная задача — *лемматизация*), т. е. выделение и сохранение корней или наиболее значимых частей слов; реализуется двумя методами — словоизменением и словообразованием [Thilagavathi, Shanmuga, 2014];
- удаление слов, не содержащих полезных сведений (или *stop-words*);
- предварительный визуальный анализ частотности словоформ текста на основе технологий векторной пространственной модели или «мешка слов» (bag of words).

В ряде случаев в качестве промежуточного этапа осуществляется процедура *очистки текста* (Text refining), преобразующая неструктурированный текст в *промежуточную форму* (intermediate form, IF) [De la Torre et al., 2005]. Различают документ-ориентированные и домен-ориентированные IF. Документ-ориентированная IF может быть спроецирована на домен-ориентированную IF путем извлечения информации, относящейся к домену.

Следующий этап обработки подготовленного текста включает в себя процедуры селекции признаков или атрибутов и селекцию шаблонов (patterns). *Селекция признаков* позволяет устранить ненужную информацию из текста и осуществляется двумя методами — фильтрацией и «упаковкой» («wrapping»), описанными, например, в [Janani, Vijayarani, 2016].

Селекция шаблонов из документ-ориентированной IF уже предполагает наличие количественных признаков и, следовательно, может быть реализована известными методами Data Mining, приведенными, например, в [Gupta et al., 2009]. В некоторых работах процедуру извлечения шаблонов, формирующих искомые знания, относят к технологии *дистилляции знаний* (*knowledge distillation*).

На завершающей стадии ТМ осуществляются *интерпретация и оценка* результата.

6. Формализация процесса извлечения знаний

В самой общей и, как следствие, малоконструктивной форме процесс извлечения знаний представляет собой реализацию некоторого оператора F_{KE}, осуществляющего преобразование множества текстовых документов $\{Doc_i,\ i = 1, \ldots, N\}$ в знания K, необходимые для решения последующих, иерархически вышестоящих задач, например задач управления (рис. 1). Формально, данное отображение можно записать в виде $F_{KE}\colon \{Doc_i,\ i = 1, \ldots, N\} \Rightarrow K$. Очевидно, что конструктивное представление оператора F_{KE} можно только в случае формализованного представления самого понятия «знание».

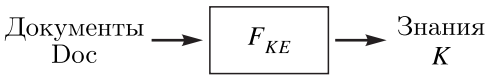

Рис. 1

Если абстрагироваться от неконкретных знаний и ограничиться решением прикладных вопросов, то знания можно представить как упорядоченные формализованные сведения, необходимые для решения конкретных прикладных задач. При этом процесс упорядочивания знаний, как правило, представим в форме некоторой классификационной схемы. В этом случае появляется возможность рассмотреть оператор извлечения знаний F_{KE} как классификационную вычислительную схему или как обобщенный оператор проверки сложных статистических гипотез.

Для иллюстрации рассмотрим вышеописанную задачу трейдинга. Входными данными является множество публикаций, формируемых финансовыми и другими аналитиками и, возможно, содержащих полезную информацию (или знания), под которыми понимается прежде всего

прогноз динамики котировок финансовых инструментов, используемых трейдером. В этом случае искомые знания можно разделить, например, на $m = 5$ классов ($C_1, \ldots, C_5$), характеризуемых нечеткими описаниями, в соответствии с которыми предполагается: C_1 — сильный отрицательный тренд, C_2 — слабый отрицательный тренд, C_3 — боковой тренд (т. е. отсутствие явного тренда), C_4 — слабый положительный тренд, C_5 — сильный положительный тренд.

Такой подход позволяет в явной форме свести задачу KE к общей задаче классификации на основе оценок классификационных признаков $\{p_j,\ j = 1, \ldots, m\}$ имеющихся текстовых документов. При использовании методов проверки статистических гипотез на основе оценок признаков формируется некоторая решающая статистика и сопоставляется с ее критическими областями $\{S_j,\ j = 1, \ldots, m\}$. Принадлежность значения статистики той или иной области $s(\widehat{p}_1, \ldots, \widehat{p}_m) \in S_{j^*}$ означает, что ожидаемая динамика котировок определяется классом C_J. При этом качество формируемого решения оценивается традиционными статистическими ошибками первого и второго рода, т. е. вероятностями принятия ошибочного решения и непринятия правильного решения. Представленная вычислительная схема представляет собой вариант классификации с учителем. В случаях, когда множество классов решений заранее не определено, используют известные вычислительные схемы кластерного анализа (классификация без учителя).

Заметим, что рассмотренная в примере вычислительная схема KE охватывает и другие задачи Text Mining, такие как автоматическая сортировка документов или задача аннотирования. В последнем случае должна быть сформирована жесткая структура аннотации. Например, первое предложение отвечает на вопрос о том, содержит ли текст информацию об интересующем трейдера инструменте. Второе предложение дает ответ о наличии явно выраженного прогноза динамики котировки. Третье предложение отвечает на вопрос о том, какой именно прогноз реализуется по мнению эксперта (например, из множества классов $C_1, \ldots, C_5$) и т. д. Каждое из предложений аннотации формируется путем последовательного применения вышеописанного подхода на основе технологий классификации. Более подробное описание процесса извлечения знаний из текстовых сообщений для принятия торгового решения приведено в работе [Мусаев, Григорьев, 2021].

7. Основные задачи Text Mining

С главной задачей TM, состоящей в *извлечении знаний из текстовых данных*, тесно связан целый ряд других важных прикладных задач, решаемых на основе тех же или близких математических технологий. Рассмотрим некоторые из них.

Обработка естественного языка [Chowdhury, 2003; Young et al., 2018; Minaee et al., 2020]. С позиции задач, решаемых средствами TM, наибольший интерес представляет задача анализа текстов, написанных на естественных языках, т. е. задача «понимания» NL текста. Одной из важнейших прикладных задач NLP является создание перспективных человеко-машинных интерфейсов, открывающее качественно новые возможности человеко-машинного симбиоза. Решение данной задачи соотносят с решением проблемы AI-полной задачи, т. е. с созданием «сильного» искусственного интеллекта.

Задачи, решаемые NLP, во многом пересекаются с задачами TM, поэтому вполне справедливо и обратное рассмотрение Text Mining, как инструмента для решения задач обработки NLP текстов. Некоторым отличием может служить акцент NLP на методах математической лингвистики, включающих в себя формализованные версии синтаксического и морфологического анализа текстов.

Реализация NLP в технологиях TM включает в себя следующие задачи [Stavrianou, Andritsos, Nicoloyannis, 2007]:

- маркировки (tagging) слов как частей речи;

- неглубокий разбор предложений (chunking) для выделения основных грамматических форм (подлежащих и сказуемых);
- выявление семантических ролей, позволяющее обнаружить параметры, ассоциированные со сказуемым в предложении;
- построение статистической модели языка, т. е. оценка вероятности возникновения последовательностей различных словоформ;
- оценивание семантической связности пар слов.

Информационный поиск (Information retrieval, IR) [Calvillo et al., 2013; Маннинг, Рагхаван, Шютце, 2011; Басипов, Демич, 2012]. Состоит в поиске информационных ресурсов, релевантных заданной теме, в БД или глобальной сети Интернет. В части, относящейся к ТМ, речь идет о поиске неструктурированной текстовой информации.

Основным отличием от классических технологий поиска, основанных на полных текстах или на метаданных, методы ТМ акцентированы на *семантическом поиске*. В этом случае текстовый документ рассматривается как объект не с точки зрения формы, а с позиции его содержания. При этом на основе NL-запроса формируется некоторый семантический образ документа, используемый в процессе поиска.

Суммаризация текста (Text Summarization) [Chen et al., 2019; Jacobs, Hoste, 2020; Белякова, Беляков, 2020]. Суммаризация текстов, по существу, представляет собой семантическое сжатие, позволяющее получить в краткой форме представление о его содержании. К суммаризации текстов можно отнести их автоматическое аннотирование и реферирование. В техническом плане различают экстрактивную и абстрактивную суммаризацию. В первом случае суммаризация осуществляется путем выявления наиболее информативных блоков (абзацы, предложения и т. п.) и их объединения в новую сжатую форму. Абстрактивный подход заключается в генерации нового текста на основе содержательного анализа и обобщения первичного документа.

Анализ тональностей (сентимент-анализ) [Araci, 2019; Sert et al., 2020]. Автоматический анализ тональности текстов предназначен для выявления общей эмоциональной окрашенности контента. При этом выявленная информация носит общий, интегральный характер и по отношению к *объекту анализа* (ОА) может носить фоновый характер.

Например, в финансовых исследованиях принято говорить об общем положительном настроении рынка в терминах «позитивный», «оптимистический», «вдохновляющий» и т. д. Имея три априорных набора словоформ в качестве обучающих данных, отвечающих трем основным тенденциям котировок (положительный, отрицательный и нейтральный), можно, используя технологии статистической классификации, сделать общие выводы о прогнозируемой динамике рыночных индексов.

Первоначально анализ тональностей, как составная часть контент-анализа, был ориентирован на гуманитарные науки — социологию и политологию. Применение методов ТМ к задачам сентимент-анализа позволило реализовать данные технологии в широком классе областей человеческой деятельности, включая экономику, финансы, бизнес и др. В более формализованных областях знаний (естественные науки, техника и технологии) эмоциональная окраска авторов, как правило, ослаблена или вообще исключена. Тем не менее данный подход может быть полезен при анализе критической инфраструктуры научных текстов. Например, можно оценить общую «окрашенность» отзывов в отношении какой-то интернет-публикации.

В Web-эпоху естественным источником для оценки мнений является сетевая информация. Проблема состоит на только в автоматическом поиске релевантных источников, но и в конструктивной оценке искомого обобщенного мнения путем обработки огромного объема субъективной, неструктурированной информации, представленной размытым и слабо выраженным

контентом. В этих условиях даже формализация задачи сентимент-анализа (или «раскопок» мнений, Opinion Mining) представляет собой непростую задачу. В качестве примера в [Liu et al., 2010] формализованное представление отдельного мнения об ОА предложено описать в виде кортежа $\langle idOA, Es, idExp, t\rangle$, где $idOA$ — идентификатор ОА, Es — оценка тональности, $idExp$ — идентификатор эксперта или субъекта анализа, t — время анализа. Эмоциональная оценка может быть выражена в нечетких лингвистических терминах типа «позитивный», «негативный» и «нейтральный». Такой подход соответствует классификации по простой интервальной шкале. Задача состоит в том, чтобы автоматически собрать множество таких кортежей, относящихся к одному ОА, и на основе полученной совокупности субъективных и нечетких суждений сформировать общее мнение с определенной объективной количественной оценкой.

Очевидно, что линейная схема оценивания может быть обобщена на случай двумерных и многомерных шкал [Bollen, Mao, Zeng, 2011; Bakshi et al., 2016]. При этом, используя набор слов, можно перейти к непрерывным шкалам тональной окрашенности, например в диапазоне $[-100, 100]$ [Thelwall et al., 2010].

Следует заметить, что для многих прикладных задач оценка общего настроения или тональности текста является недостаточной информацией. В частности, по общему заключению о положительном настрое экспертов по отношению к динамике рынков крайне трудно оценить динамику котировок конкретного инструмента. Поэтому в прикладных исследованиях применяется совместное решение таких задач, как *извлечение сущностей* (entity extraction), *моделирование тематики* (topic modelling) и собственно *сентимент-анализ* [Sun, Lachanski, Fabozzi, 2016; Xing, Hoang, Vo, 2020].

В качестве примера рассмотрим задачу трейдинга на основе валютной пары USDRUB. Если тональность рассматриваемых аналитических отчетов указывает на активизацию промышленной деятельности в мире, это означает рост спроса на нефть и, как следствие, рост котировок рубля в валютных парах.

На практике часто возникает задача выявления тональности текстов относительно конкретной сущности, например мнение о USDRUB. Для этого используют методы *целевой тональности* (targeted sentiment) [Jiang et al., 2011], когда с помощью синтаксического и семантического аннотирования текста выделяется контекст требуемой сущности.

8. Категоризация текстов в задачах извлечения знаний

Одной из наиболее распространенных форм IE является категоризация текстов, т. е. путем их отнесения к одной из априори заданных категорий или классов. По мере детализации категорий получаем последовательное раскрытие семантического контента документа. В этом случае в зависимости от априорного наличия или отсутствия множества классов и их характеризаций используются методы классификации и кластеризации.

Классификация текста [Nasa, 2012; Jacobs, Hoste, 2020]. Задача состоит в сортировке текстовых документов путем отнесения к одному из заданных семантических классов. Классификация является важнейшим элементом когнитивного восприятия реальности, позволяющим человеку или искусственному интеллекту формировать управляющие решения и взаимодействовать с окружающей средой. Классификация состоит в оценке контента текстового документа, достаточной для соотнесения содержания документа с одним из априорно выбранных классов.

При статистическом подходе в качестве классификационных признаков обычно используются частоты словоформ. Очевидно, что в этом случае предполагается наличие большой обучающей выборки текстов, позволяющей сформировать тезаурус словоформ, определить наиболее вероятные словоформы различной длины, а также установить связи между словоформами. Кроме того, категоризация обычно содержит метод ранжирования документов по величине оценки вероятности принадлежности их контента к тому или иному классу.

Задачи классификации основаны на технологиях *машинного обучения с учителем* (supervised technique), под которым понимаются априори заданные наборы шаблонов входных и выходных данных, позволяющие построить дискриминационные модели. При этом широко используются техники на основе деревьев решений, ассоциативной классификации, графовых моделей терминов, байесовских классификаторов (в том числе метод наивного Байеса), k-ближайших соседей, опорных векторов (SVM), генетических алгоритмов, нечетких корреляций, алгоритма Роккио и др. [Gupta et al., 2009; Atika, Ali, Ahmer, 2009; Stavrianou, Andritsos, Nicoloyannis, 2007]. В ряде случаев при решении классификации текстов используются *искусственные нейронные сети* (ИНН).

Кластеризация текста [Kumar, Bhatia, 2013; Nalini, Sheela, 2014]. В случае отсутствия априорной информации о классах рассматриваемых документов используются методы кластеризации, связанные с технологиями *машинного обучения без учителя* (unsupervised technique). Кластеризация осуществляет сортировку документов одновременно с формированием классов или кластеров, с которыми эти документы соотносятся. Данная технология широко используется в задачах глубокого обучения (Deep Learning). При этом один и тот же документ может быть отнесен одновременно к нескольким классам, что снижает вероятность потери нужной информации.

Обычно в процессе кластеризации каждому документу сопоставляется вектор тем (topics), определяющий весовую меру соответствия каждому кластеру. Различные методы кластер-анализа могут приводить к различным классификационным группам или категориям.

Обычно методы кластеризации делят на иерархические, использующие древовидную структуру или дендограммы, и неиерархические. В иерархической структуре каждый кластер представляет собой узел, который разделяется на нижестоящие кластеры. Такой подход позволяет реализовать последовательную детализацию текстов по их контенту. При этом иерархия может строиться как снизу вверх (агломерация), т. е. путем слияния кластеров, так и, наоборот, сверху вниз путем их разделения. В неиерархических схемах все множество имеющихся текстов делится на заданное число пересекающихся или непересекающихся кластеров. Для взаимоисключающих категорий используются методы секционирования или, значительно реже, методы маркировки.

Кластеризация текстов осуществляется на основе технологий, основанных на иерархическом методе, методе секционирования [Ghosh, Roy, Bandyopadhyay, 2012], методе k-средних [Nasa, 2012], кластеризации на основе относительности слов [Gupta et al., 2009], техники внутриклассового подобия (intra-cluster similarity, IST) [Steinbach, Karypis, Kumar, 2000], алгоритма на основе плотности распределения [Berkhin, 2002] и др.

9. Методологические аспекты Text Mining

Общим вопросом всех задач и технологий TM является выбор типа словоформ и их производных, используемых для распознавания контента в последовательностях символов NL. Выбор эталонных форм, по существу, определяет дальнейшее развитие методологической и алгоритмической базы, используемой для решения поставленной задачи извлечения знаний. Рассмотрим основные подходы на примере задачи информационного поиска.

Первичную классификацию поисковых методов можно сформировать на основе двух классов: статистический и семантический. В первом случае поиск осуществляется на основе анализа частот возникновения в тексте сочетаний символов, приведенных в запросе пользователя. Во втором случае акцент сделан на смысловой стороне этого запроса.

В техническом плане IR можно систематизировать по типу словоформ и их сочетаний, используемых в процессе поиска.

Поиск на основе словоформ (Term Based Method) [Salton, Buckley, 1988]. Данный тип поиска наиболее распространен и является основой поиска информации по ключевым словам.

Очевидно, что повышение качества селекции документов при таком подходе будет связано с частотным анализом словоформ. В этом случае в качестве распознающих признаков выступает частота, т. е. оценка вероятности появления заданной словоформы в тексте, содержащем нужный контент.

Такой подход требует *обучающего полигона* (Data Set), в котором множество документов априори разделены экспертами по классам распознаваемого контента. Автоматическое формирование классов, т. е. кластеризация данных, обычно является менее эффективным средством, так как весьма вероятны случайные варианты группировок, не связанные с контентом.

Технологии распознавания, основанные на словоформах (или терминах), в меньшей степени подвержены различным вариантам неоднозначностей, имеющих место в NL. Наиболее характерными для таких языковых особенностей являются синонимы и полисемия, т. е. семантическая многовариантность слов.

Поиск на основе словосочетаний (Phrase Based Method) [Ahonen et al., 1998]. Достаточно понятно, что формирование запроса из нескольких семантически связанных словоформ (фраз) повышает качество селекции, однако в этом случае существенно усиливается проблема «малой вероятности». Собственно говоря, она имеет место и для поиска на основе отдельных словоформ, но в этом случае эта проблема радикально усиливается. По существу, частотные оценки оказываются соизмеримыми с частотами шумовых словоформ.

Отказ от статистического подхода вновь возвращает методику к проблеме некачественной селекции, т. е. ошибкам первого и второго рода: селекция текстов, не содержащих релевантной информации, и отбраковка семантически значимых текстов.

Поиск на основе шаблонов (patterns) [Lodhi et al., 2002; Wu et al., 2004]. На основе семантического анализа обучающей информации формируются синтаксические шаблоны, характерные для заданной тематической области. В качестве шаблонов могут выступать отдельные словоформы и фразы. Однако данный подход позволяет сформировать достаточно необычные сочетания символов (шаблоны), существенно повышающие качество семантической селекции. Иными словами, поисковая система сначала должна по обычному текстовому запросу грубо оценить семантический класс (или кластер) и на его основе реструктурировать запрос в соответствии со сформированными для заданного класса шаблонами. Такой подход является существенным продвижением к семантическому поиску.

Поиск шаблонов в БД является одной из центральных задач Data Mining, в рамках которой разработаны методы выявления:

- ассоциаций (mining associations) [Ghafari, Tjortjis, 2019; Hipp, Güntzer, Nakhaeizadeh, 2000],
- правил (rule mining) [Rajak, Gupta, 2008],
- частотных подмножеств (frequent item set mining) [Borgelt, 2012; Thirumuruganathan et al., 2014; Zeng, Naughton, Cai, 2012],
- последовательных шаблонов (последовательный анализ шаблонов, sequential pattern mining) [Hosseininasab, van Hoeve, Cire, 2019; Aloysius, Binu, 2013; Tax et al., 2016],
- скрытых закономерностей (шаблонов) (closed pattern mining) [Fradkin, Moerchen, 2010; Kaytoue, Kuznetsov, Napoli, 2011; Buzmakov, Kuznetsov, Napoli, 2015].

В то же время обеспечение эффективности IR на основе шаблонов сталкивается с известным противоречием между проблемой низких частот и вероятностью ошибок второго рода — сбора информации, не содержащей полезного контента. Оптимум ищется последовательной модификацией шаблонов, их расширениями и изменениями.

Регулярные выражения (Regular Expressions) представляют собой реализацию поиска точных вхождений подстроки в тексте, заданной определенным шаблоном. Такая технология достаточно оперативна, поскольку действуют за один проход путем построения конечных автоматов для заданного шаблона. Применение такого подхода полезно на стадиях очистки текста: токенизации, удаления стоп-слов, пунктуации и идентификаторов, а также при выделении требуемых данных из веб-страниц. Однако регулярные выражения могут выдавать лишние данные, совпадающие по шаблону с целевыми, поэтому имеет смысл применять их совместно с другими средствами NLP и машинного обучения. Например, в [Zhong et al., 2018] предлагается подавать извлеченные регулярными выражениями данные на вход ИНН для ее предварительного обучения. Затем пользователь размечает небольшое количество эталонных данных, чтобы окончательно настроить параметры ИНН для более точного извлечения требуемых данных.

Повысить качество извлечения информации регулярными выражениями можно, например, с помощью фреймворка *TokensRegex*, входящего в пакет *Stanford CoreNLP*. Данный подход предлагает проводить поиск шаблона не только по тексту, но и по результатам его NLP-обработки: токенам, частям речи (POS) и именованным сущностям (NER). Таким образом, возникает возможность формирования связанных пар, например, правила извлечения пары «сотрудник–компания»:

$$(([\{ner: \text{“PERSON”}\}]+)\ (/works/\ /for/)|(/is/\ /employed/\ /at|by/))\ ([\{ner: \text{“ORGANIZATION”}\}]+)$$

Дальнейшее улучшение такого подхода возможно путем использования в регулярных выражениях информации о зависимостях слов (dependency graph). Например, библиотека *Semgrex*, также построенная на базе Stanford CoreNLP, представляет дерево зависимостей в виде матрицы «атрибут–значение» и применения регулярных выражений над подмножествами атрибутов [Chambers et al., 2007]. Кроме того, в библиотеке поддерживаются отношения зависимости и предшествования между узлами дерева. Так, например, паттерн «$\{word: is\} >nsubj\ \{\}$ = субъект $>obj\ \{\}$ = объект» выделит из предложения Musk is a director of Endeavor Group Holdings субъект Musk и объект director в отношении действия is.

Поиск на основе концептов [Lee, 2009; Ghorbani et al., 2019; Prokasheva, Onishchenko, Gurov, 2013]. Данный подход предполагает предварительное выявление концепта документа, т. е. сжатого семантического шаблона, отражающего его содержание. При этом шаблон должен соответствовать априори выбранному признаковому пространству, допускающему упорядоченный поиск традиционными средствами Data Mining. Модель IR на основе концептов включает в себя следующие этапы:

- компонентный анализ семантической структуры документа;
- построение концептуального онтологического графа, позволяющего описать семантическую структуру текста;
- формирование топовых концептов, основанных на первых 2-3 компонентах и используемых в качестве распознающих признаков в векторном пространстве моделей.

Данный подход позволяет провести дискриминацию на семантическом уровне между содержательно значимыми и незначимыми текстами, что в свою очередь повышает вероятность селекции релевантной информации.

10. Извлечение знаний из текстовых документов: обзорные публикации

Прежде всего рассмотрим работы, непосредственно связанные с главной проблемой ТМ, т. е. с задачей извлечения знаний (Information Extraction, IE).

В обзорной работе [Thilagavathi, Shanmuga, 2014] приведен набор технических решений задачи IE, причем приоритет отдается лексическому анализу. Предложен эффективный алгоритм частотного выявления шаблонов с использованием техники кластеризации и метода k-ближайших соседей. В работе [Atika, Ali, Ahmer, 2009] рассмотрены вопросы категоризации текстов и извлечения информации на основе технологий Data Mining. При этом используется частотный подход, при котором оценки вероятности появления словоформ формируются на основе доменных словарей.

В статье [Anshika, Udayan, 2013] сделан акцент на лексическом и синтаксическом анализе. Рассмотрены основанные на правилах (rule-based, RB) системы IE с примерами из химических технологий. Обзорная статья [Allahyari et al., 2017] рассматривает особенности применения технологий Data Mining и обнаружения знаний в БД (knowledge discovery in databases, KDD) в задачах ТМ с примерами в области распознавания биомедицинских текстов. В работе [Umajancy, Thanamani, 2013] указывается на важность интеграции базы знаний предметной области с механизмом интеллектуального анализа текста для повышения эффективности ТМ, особенно на этапах поиска и извлечения информации. Особое внимание уделяется технологиям для селекции текстов и других приложений.

Во многих работах задача IE рассматривается одновременно с вопросами информационного поиска. В частности, в [Atika, Ali, Ahmer, 2009] для формирования признакового пространства используются словари стемминговых доменов, а в обзоре [Sagayam, Srinivasan, Roshni, 2012] приведена техника индексации с использованием булевских IR-моделей и *модели векторного пространства* (Vector Space Model, VSM). При этом сложность возникает в результате несоответствия понятий релевантности и подобия текстовых документов, а также в результате нечеткой экспликации формируемых запросов на NL. Нейросетевой подход к задаче IR и NLP представлен во многих работах, например в [Collobert, Weston, 2008], в которой представлена задача применения односверточной искусственной нейронной сети к задаче обработки текстов.

11. Технологии NLP и математическая лингвистика в задачах извлечения знаний

При решении задач IE важное значение имеет часть NLP [Agarwal, 2019; Ding et al., 2016; Khurana et al., 2017], связанная с пониманием естественных языков, т. е. с лингвистическими технологиями. В связи с этим ТМ пересекается с такими разделами лингвистики, как морфология (технологии формирования и анализа слов), синтаксис (технологии формирования и анализа предложений) и, естественно, семантика, т. е. с содержательным анализом. Рассмотрим эти вопросы немного подробнее.

Морфологический анализ, позволяющий выделять морфемы или наименьшие значимые языковые единицы. К морфемам относятся значимая часть слова, или корень слова, и его вспомогательная часть, или аффикс. Аффиксы подразделяются на префиксы и постфиксы, которые располагаются соответственно до и после корня. В русском языке префиксы — это приставки, а постфиксы — это суффиксы и окончания. Морфологический анализ состоит в сопоставлении словоформ (слов текста) и их лексем (словарных форм). Таким образом, морфологический анализ позволяет перейти от множества словесных форм к унифицированной форме, к его корню, что существенно упрощает задачу компьютерного IE.

Характер морфологического анализа в значительной степени зависит от анализируемого языка. В некоторых языках отдельные слова (используемые как глаголы) содержат всю информацию о времени, роде и числе и так далее. В других языках эта информация может быть передана через несколько слов в предложении. Например, в англоязычном предложении He will have wrote this essay by Monday сложная информация о времени передается через вспомогательные глаголы will и have. В русскоязычном предложении «Он напишет сочинение к понедельнику» информация о времени передается с помощью аффикса глагола «напишет».

С точки зрения компьютерной обработки текста морфологический анализ позволяет осуществить POS-тегирование, т. е. каждому слову текста сопоставляются часть речи и набор морфологических характеристик.

Синтаксический анализ позволяет выявить структуру синтаксических отношений между словами предложения. При этом синтаксический анализатор использует словарь определений слов (лексикон) и набор синтаксических правил (контекстно-свободную грамматику). Простой лексикон содержит только синтаксическую категорию каждого слова, простая грамматика описывает правила, которые указывают только на то, как синтаксические категории могут быть объединены для формирования фраз разных типов. Результаты синтаксического анализа применяются затем при анализе семантики, поэтому особенно важно определить наилучшую структуру предложения. Однако неоднозначность естественного языка приводит к тому, что к одному и тому же предложению можно составить несколько синтаксических схем и добавление новых правил в грамматику практически не дает результата [Jurafsky, Martin, 2020]. Современные алгоритмы синтаксического анализа построены на вероятностных правилах, что может приводить к построению неоптимальной схемы предложения и затруднить дальнейшее решение поставленной задачи. На практике такую неоднозначность пытаются устранить составлением «банков деревьев» (treebanks), в которых значительное количество предложений размечено лингвистами. Для английского языка доступно порядка 30 банков, в то время как для русского — только 5 (GSD, PUT и т. п.). Таким образом, чтобы снизить вероятностную ошибку синтаксического разбора, необходимо выбрать релевантный целевому корпусу документов «банк деревьев».

Семантический и прагматический анализ состоит в понимании высказывания. Понимание представляет собой крайне сложный, плохо формализуемый процесс семантической интерпретации в контексте высказывания, который зависит от результатов предыдущих этапов NLP, от лексической информации, контекста и здравого смысла.

Проектирование семантического интерпретатора предполагает решение тех же проблем, с которыми приходится сталкиваться при построении синтаксического анализатора, в частности проблемы (семантической) двусмысленности. В простейшем случае работу интерпретатора можно свести к задаче классификации, т. е. распознавания предполагаемой семантической интерпретации высказывания в конкретном контексте среди множества возможных интерпретаций этого предложения. При этом остается открытым вопрос о том, какой именно должна быть конечная интерпретация высказывания. Прикладные системы NLP, как правило, используют семантические представления, предназначенные для конкретной предметной области.

Семантический анализ связывает смысл с изолированными высказываниями (предложениями), прагматический анализ интерпретирует результаты семантического анализа с точки зрения конкретного контекста. В некоторых случаях прагматический анализ находит соответствие реальным объектам или событиям, которые существуют в данном контексте, со ссылками на объекты, полученными в ходе семантического анализа.

К общим задачам NLP и TM относят решение таких вопросов, как:

- автоматическая суммаризация (реферирование, аннотирование), т. е. краткое изложение содержания заданного набора текстов;

- машинный перевод с одного NL на другой;
- автоматический дискурс-анализ, позволяющий выявлять спорные вопросы в связанных текстах;
- категоризация текстов и др.

В то же время NLP присущи характерные для этой области задачи, на которых могут основываться предобработка текстов и извлечение знаний:

- морфологическая сегментация текстов, позволяющая разделять слова на индивидуальные морфемы, а затем формировать классы морфем;
- кореферирование, т. е. выявление различных слов, относящихся к одному объекту;
- формирование терминологий, выделение ключевых слов и коллокаций;
- распознавание *именованных сущностей* (Named Entity Recognition, NER), состоящее в выявлении словоформ, обозначающих предмет или явления определенной категории;
- разметка частей речи (part-of-speech tagging, POS-тегирование) — определение частей речи в предложении для каждого слова;
- *оптическое распознавание символов* (optical character recognition, OCR) — преобразование изображений текста в текстовые данные, использующиеся для машинного представления символов и т. п.

Извлечение признаков состоит в [Sesen, Romahi, Li, 2019] трансляции предобработанного текста в векторное цифровое представление, подходящее для моделей машинного вывода. Простейшим представлением является «мешок слов» (bag-of-words) — неупорядоченное множество уникальных слов из предобработанного текста, часто ограниченное по размеру наиболее часто встречающимися словами. Такое представление не учитывает ни синтаксиса, ни контекста слова, однако позволяет рассчитать его частотные характеристики в документе. Частично контекст слова возможно учесть его заменой в «мешке слов» на последовательность N соседних слов (N-граммы). Большие значения N позволяют учесть больше информации контекста, однако из-за вычислительных ограничений на практике ограничиваются биграммами и триграммами, которые формируют пары и тройки соседних слов соответственно.

В то же время нельзя полагаться на то, что наиболее распространенные слова в данном документе будут наиболее важными с точки зрения поставленной задачи, например предлоги и местоимения. Метод Tf-idf (term frequency – inverted document frequency) нормализует частоту слова в данном документе с учетом частоты его появления в других документах, что дает необходимую информацию о важности слова в рассматриваемом корпусе документов.

Модели машинного обучения, которые осуществляют задачи классификации, регрессии и кластеризации, требуют подавать на вход данные в виде векторов чисел. Самый простой способ — представить предобработанный текст в виде вектора из индексов слов в некоем общем словаре. Однако таким образом не получится учитывать информацию о синонимах. Модели «переноса знаний» (transfer learning) обучаются на значительном корпусе текстовых документов (Word2Vec и GloVe) и представляют слово как вектор в общем пространстве таким образом, чтобы векторы похожих слов находились рядом.

В результате модель машинного вывода выдает свое решение, которое требуется интерпретировать и качественно оценить. В случае когда на выходе имеем вероятностную оценку

(например, в задаче классификации) решения, то либо оно автоматически принимается при превышении некоего вероятностного порога, либо решение принимает эксперт с учетом всех за и против по совокупности вероятностных решений на выходе [Sesen, Romahi, Li, 2019].

Последовательность применения указанных методов в конвейере NLP-обработки представлена на рис. 2.

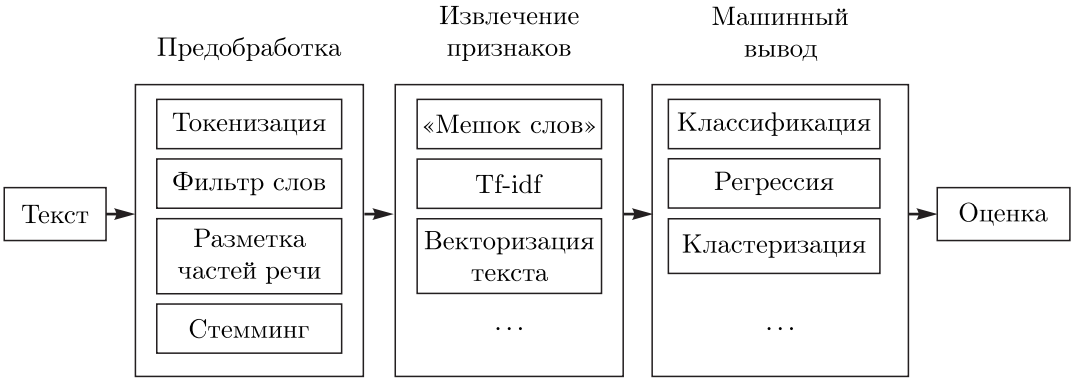

Рис. 2

В [Xing, Cambria, Welsch, 2018] представлен обзор конкретных NLP-конвейеров и систем машинного вывода для построения метасистем принятия торговых решений на финансовых рынках. Их тестирование на исторических данных показало перспективность учета текстовой информации для трейдинга. Вместе с тем NLP тесно связан с задачами математической или компьютерной лингвистики, центральной задачей которой является создание математических моделей естественных языков. Многие направления математической лингвистики совпадают с задачами NLP, но имеются и некоторые специфические вопросы, такие как создание и использование электронных корпусов текстов (корпусная лингвистика), создание электронных словарей, онтологий и тезаурусов, решение задач по автоматическому переводу текстов.

12. Программные инструменты Text Mining

Рассмотрим некоторые наиболее распространенные программные средства и комплексы, используемые для решения задач TM.

GATE (General Architecture for Text Engineering, обобщенная архитектура для обработки текста) — программный комплекс для NLP с открытым кодом, впервые разработанный в Университете Шеффилда в 1996 г. Написан на платформе Java и с помощью компонентных интерфейсов, определенных в XML, имеет возможность интегрироваться со сторонними программными средствами, такими как системы извлечения информации, статистические пакеты и фреймворки машинного обучения. Систему можно использовать как библиотеку или как отдельное приложение.

GATE позволяет частично структурировать текст путем автоматического формирования и добавления аннотаций к сегментам текста. Кроме основной задачи IE, реализуемой в основном через создание правил вручную (Java Annotation Pattern Engine, JAPE) или средствами машинного обучения над аннотациями, комплекс позволяет также анализировать потоки сообщений и работать с онтологиями (здесь: *иерархическими описаниями информации из конкретных областей знаний*).

Конвейерная (pipeline) архитектура с унифицированными интерфейсами позволяет формировать требуемую структуру из разных компонентов для каждой конкретной задачи TM. Компоненты GATE разделены на три группы ресурсов: языковые (language), процессинговые (processing) и средства визуализации. Качество работы процессингового ресурса может быть

оценено с помощью нескольких встроенных метрик. В качестве входа поддерживаются все основные текстовые форматы (Text, HTML, SGML, XML, RTF, Email, PDF и др.) и все наиболее распространенные NL (английский, русский, испанский, китайский, арабский, французский, немецкий, хинди и т. д.).

Система содержит интеллектуальную подсистему ANNIE (A Nearly-New Information Extraction System), позволяющую решать достаточно сложные задачи предобработки текстов, в том числе автоматическую морфологическую разметку, анализ кореференции, т. е. установление референциональных тождеств между именами, извлечение именованных сущностей и др. Кроме мощных готовых инструментов и словарей (gazetter), в GATE доступны дополнительные плагины с альтернативной реализацией и новыми функциями.

На практике автоматическое аннотирование документов корректируют ручной обработкой и последующим переобучением процессинговых моделей. Проект GATE Teamware поддерживает через веб-интерфейс этот процесс и организует совместную работу аннотаторов, кураторов, проверяющих их работу, координирующих менеджеров и системных администраторов.

Развитие NLP-методов и появление большого количества инструментов вызвало необходимость в принятии стандарта **UIMA** (Unstructured Information Management Application), который обеспечил возможность декомпозировать NLP-задачи по компонентам и выстраивать их в конвейер, аналогично GATE. Компания IBM представила реализацию UIMA, которая затем была передана сообществу Apache Foundation в 2006 году. Графический интерфейс выполнен в кроссплатформенной среде разработки Eclipse. Отличительными чертами системы является концентрация на производительности и масштабируемости. Ее компоненты могут быть написаны на языках Java или C++, а затем распределены в виде веб-сервисов. Согласованность интерфейсов обеспечена декларативным описанием в XML типов входных и выходных данных каждого компонента. В то же время стандарт UIMA не накладывает никаких ограничений на реализацию компонентов, что дает гибкость и производительность всей системы, однако привносит дополнительную сложность по сравнению с разработкой в GATE.

В свою очередь, коммерческий продукт **RapidMiner** предлагает упрощенный подход к подготовке и сбору данных, машинному обучению и предиктивной аналитике. Весь процесс — от сбора данных до оценки модели — заключается в визуальном проектировании суперпозиции операторов, которые преобразуют и обрабатывают данные. Исходными источниками могут быть: веб-страницы, текстовые документы, аудио- или видео-файлы. Доступно более 400 встроенных операторов, а также возможности по созданию новых на языках WEKA, Java, Python и R. Визуальная среда разработки предиктивных моделей RapidMiner Studio содержит все требуемые средства и расширяема с помощью сторонних плагинов. Кроме того, поддерживается работа через командную строку и Java API. Инструмент подходит для работы с большими данными, делегируя управление и расчеты на сервер RapidMiner Server, в облако (Amazon AWS или Microsoft Azure) или кластер (Radoop). Первоначальная версия RapidMiner была выпущена под свободной лицензией в 2001 году Техническим университетом Дортмунда.

Популярной альтернативой по анализу данных «без программирования» в академических кругах является свободный инструмент **WEKA** (Waikato Environment for Knowledge Analysis), развивающийся с 1993 года. Аналогично RapidMiner он имеет графический интерфейс для проектирования и исследования моделей. Доступны компоненты для традиционных задач: предварительной обработки данных, кластеризации и классификации, регрессии, глубокого обучения, визуализации, отбора признаков и т. п. Наиболее подходит для быстрого прототипирования моделей, после чего их можно интегрировать в Java-, Python- и R-программы. По сравнению с RapidMiner имеет более ограниченные возможности визуализации и отладки построенных моделей, а также поддерживается работа с большими данными на кластерах только через Apache Spark.

Konstanz Information Miner (**KNIME**) также популярен в бизнес-среде и, в частности, финансовом секторе. Следуя концепции no-code, он предлагает в рамках среды Eclipse свыше 1000 готовых аналитических компонентов для проведения уни- и мультивариантного анализа, статистических тестов, анализа временных серий, веб-аналитики, анализа социальных сетей и т. п. Создание новых компонентов производится на языках R и WEKA. Поддерживается полный комплекс работы с данными по принципу «*извлечь, преобразовать, загрузить*» (ETL): фильтрация, агрегация, объединение данных в локальной или распределенной среде, очистка данных путем нормализации, преобразования к другим типам и заполнения пропущенных признаков и т. д. Инновационной особенностью системы можно считать KNIME Testing Framework — автоматизированное тестирование развернутых моделей, что делает его пригодным для полного цикла — от экспериментов до внедрения. Для совместной работы и процесса автоматизации предлагается платная версия KNIME Server. Аналогично вышеупомянутым системам кросс-платформенный KNIME написан на Java и предложен в начальной версии учеными Konstanz University в 2006 году.

Проект с открытым кодом **Orange**, разработанный в Университете Любляны (Словения), интересен в первую очередь для Python-сообщества. Он представляет собой графический фреймворк, в котором можно применять блоки Data, Visualize, Classify, Regression, Evaluate, Associate и Unsupervised. Ядро системы написано на C++, ее применение организовано в визуальном редакторе, а расширяемость — через Python. Таким образом обеспечивается баланс производительности и удобства применения. Orange имеет активное сообщество, которое по мере появления более мощных технологий ТМ интегрирует их в систему. Аналогично WEKA проект Orange имеет длительную историю, начиная с 1996 года.

Отдельный набор программных библиотек ориентирован на связанные с ТМ задачи NLP. Пакеты свободных библиотек **NLTK** (Python), **OpenNLP**, **Stanford CoreNLP** (Java) представляют инфраструктуру для обработки текстов с помощью средств машинного обучения. Типичными задачами, которые решаются с помощью этих инструментов, являются: предобработка, очистка и векторизация текста, синтаксический анализ и POS-тегирование, лемматизация, NER, анализ зависимостей и кореференции. С каждым инструментом поставляются предварительно обученные модели для различных языков, включая русский. Дополнительно решаются задачи классификации текстов, тематического моделирования, суммаризации текста и анализа тональности. Отметим, что эти библиотеки могут быть основой для соответствующих компонентов графических приложений по ТМ, которые обсуждались выше, однако NLTK имеет собственные средства визуализации в рамках консоли.

Для структурирования и выделения информации из «плоских» текстов на русском языке существует ряд специализированных инструментов. Компанией «Яндекс» разработан **Томита-парсер**, позволяющий, аналогично JAPE, выявлять факты путем определения правил в виде расширяемых контекстно свободных грамматик и создания словаря ключевых слов. Терминалами в таких грамматиках являются части речи, леммы слов, регулярные выражения. В словаре задаются слова, которые могут участвовать в описываемых фактах. Разработчики применяют свой инструмент для обогащения результатов поисковых запросов, однако не предоставляют сформулированные грамматики и словари. Проект является открытым и распространяется в виде консольного приложения. Сходную функциональность имеет открытая библиотека для Python — **Natasha**. В то же время наиболее высокое качество и количество извлекаемых сущностей для русского языка демонстрируют коммерческие библиотеки **ABBYY FlexiCapture** и **RCO Fact Extractor**.

Gensim — еще одна высокопроизводительная библиотека для Python, сфокусированная на задачах векторизации текстов, индексирования документов и семантического поиска, а также *тематического моделирования* (topic modeling). Отличием от других библиотек является возмож-

ность обработки больших документов через итераторы, что снимает ограничение на их полную загрузку в память.

В настоящее время наиболее качественные результаты при NLP-анализе показывают такие модели обработки текста, как *трансформеры* (**BERT** [Devlin et al., 2018], **GPT-2** [Radford et al., 2019] и **GPT-3** [Brown et al., 2020]). Прогресс в вычислительных мощностях и развитии архитектур глубокого обучения позволил в 2020 году создать для GPT-3 нейронную сеть с 175 млрд параметров и обработать колоссальные объемы англоязычного текста (45 TB). В 2021 году этот рекорд был побит трансформером MT-NLG от компаний Microsoft и Nvidia с 530 млрд параметрами и обучающими данными в размере 270 млрд слов. Обученную на общем множестве данных модель, учитывающую контекст, можно затем перенести (transfer learning) на более узкие области: медицину [Kalyan, Sangeetha, 2021], финансы [Liu et al., 2020], туризм [Chantrapornchai, Tunsakul, 2021] и т. п. Библиотека **SpaCy** для Python предназначена для решения стандартных задач NLP-анализа подобными современными методами глубокого обучения.

Полученные в результате NLP-анализа аннотированные признаки документов для решения задач ТМ подаются на вход инструментам машинного обучения, таким как **Keras**, **FastAI**, **PyTorch** и др. Получается, что программисту требуется хорошо ориентироваться в нескольких довольно сложных инструментах. Однако для конкретных задач существуют обобщающие библиотеки. Например, **NLP Architect** и **DeepPavlov** применяются для создания диалоговых и вопросно-ответных чат-ботов. В них есть все необходимые компоненты лингвистического, синтаксического и семантического анализа текста, от проверки орфографии до извлечения намерения пользователя (intent extraction), а также системы понимания смысла текста. Таким образом, в рамках одного инструмента можно реализовать диалоговую систему, подводящую пользователя к какой-то цели или предоставляющую справочные ответы на свободно сформулированные вопросы.

Лидеры IT-индустрии также развивают облачные продукты для области NLP и смыслового понимания текстов (Natural Language Understanding, NLU): **Google Cloud Natural Language**, **IBM Watson NLP/NLU** и **Microsoft Text Analytics**. В первую очередь эти инструменты удобны неискушенному в программировании пользователю тем, что предоставляют «из коробки» средства для анализа своих документов на основе предварительно обученных моделей, а также возможность предоставить свои размеченные данные для улучшения результата. Поддерживается множество входных форматов данных: простые тексты, документы (в том числе и графическом виде), социальные сети и новостные ресурсы, аудио- и видео-файлы. Дополнительным преимуществом являются скорость получения результата и возможность развернуть систему для обработки приходящих данных в режиме «онлайн». В то же время в этих системах невозможно повлиять на качество результатов моделирования путем их настройки, допустимо только увеличивать размер предоставляемых обучающих данных.

Обобщенные характеристики упомянутых инструментов в соответствии с набором, предложенным в [Ratra, Gulia, 2021; Bartschat, Reischl, Mikut, 2019], приведены в приложении. Согласно [Bartschat, Reischl, Mikut, 2019], перечисленные инструменты можно разделить по типам DMS (Data Mining Suites — пакеты с графическим интерфейсом по извлечению данных), SOL (Solutions — среды разработки, поддерживающие полный процесс по задачам ТМ), LIB (Libraries — библиотеки функций, которые могут быть интегрированы в стороннее ПО) и BIN (Binaries — скомпилированные исполняемые приложения). Таким образом, продукты типов DMS и SOL будут удобны для бизнес-аналитиков, в то время как LIB и BIN — для программистов.

13. Заключение

Задача извлечения знаний из текстовых сообщений представляет собой одну из актуальных составляющих общей проблемы построения систем ИИ. Актуальность данной задачи связана

с накоплением огромного объема плохо структурированных данных, содержащих важные, порою латентные знания, необходимые для построения эффективных систем управления сложными объектами различной природы. Данная проблема непосредственно связана с известным трендом в области информационных технологий, получившим наименование Big Data.

С другой стороны, технологии Text Mining по своему назначению можно рассматривать как составляющую часть Data Mining, другого направления развития ИТ, ориентированного на извлечение знаний из баз данных различной природы.

В зарубежной периодике регулярно возникают публикации, посвященные как оригинальным исследованиям в области ТМ, так и обобщающие статьи обзорного типа. В русскоязычной периодике данные работы представлены очень ограничено. Восполнению этого пробела и посвящена настоящая статья.

Список литературы (References)

Басипов А. А., Демич О. В. Семантический поиск: проблемы и технологии // Вестник Астраханского государственного технического университета. Сер. Управление, вычислительная техника и информатика. — 2012. — № 1.

Basipov A. A., Demich O. V. Semanticheskii poisk: problemy i tekhnologii [Semantic Search: Problems and Technologies] // Vestnik Astrakhanskogo gosudarstvennogo tekhnicheskogo universiteta. Ser. Upravlenie, vychislitel'naya tekhnika i informatika. — 2012. — No. 1 (in Russian).

Барсегян А., Куприянов М., Степаненко В., Холод И. Технологии анализа данных: Data Mining, Text Mining, Visual Mining, OLAP. — 2-е изд. — БХВ-Петербург, 2008.

Barsegyan A., Kupriyanov M., Stepanenko V., Kholod I. Tekhnologii analiza dannykh: Data Mining, Text Mining, Visual Mining, OLAP [Data analysis technologies: Data Mining, Text Mining, Visual Mining, OLAP]. — 2nd edition. — BHV, 2008 (in Russian).

Белякова А. Ю., Беляков Ю. Д. Обзор задачи автоматической суммаризации текста // Инженерный вестник Дона. — 2020. — № 10. — С. 142–159. — URL: http://www.ivdon.ru/uploads/article/pdf/IVD_33__10_belyakova_belyakov.pdf_f5b35cc7eb.pdf (дата обращения: 14.04.2021).

Belyakova A. Yu., Belyakov Yu. D. Obzor zadachi avtomaticheskoi summarizatsii teksta [Overview of text summarization methods] // Inzhenernyi vestnik Dona [Engineering journal of Don]. — 2020. — No. 10. — P. 142–159. — Available at: http://www.ivdon.ru/uploads/article/pdf/IVD_33__10_belyakova_belyakov.pdf_f5b35cc7eb.pdf (accessed: 14.04.2021).

Винер Н. Кибернетика, или управление и связь в животном и машине; Кибернетика и общество. — 2-е издание. — М.: Наука; Главная редакция изданий для зарубежных стран, 1983. — 344 с.

Wiener N. Cybernetics or Control and Communication in the Animal and the Machine. — MIT press, 2019. (Russ. ed.: *Viner N.* Kibernetika, ili upravlenie i svyaz' v zhivotnom i mashine; Kibernetika i obshchestvo. — 2-e izdanie. — M.: Nauka; Glavnaya redaktsiya izdanii dlya zarubezhnykh stran, 1983. — 344 p.)

Гордиенко Е. П., Паненко Н. С. Современные технологии обработки и анализа больших данных в научных исследованиях // Актуальные проблемы железнодорожного транспорта. — 2018. — С. 44–48.

Gordienko E. P., Panenko N. S. Sovremenny'e texnologii obrabotki i analiza bol'shix danny'x v nauchny'x issledovaniyax [Modern technologies for processing and analyzing big data in scientific research] // Aktual'ny'e problemy' zheleznodorozhnogo transporta. — 2018. — P. 44–48 (in Russian).

ГОСТ 7.0-99 Информационно-библиотечная деятельность, библиография. Термины и определения. — Дата введ. 2000-07-01.

GOST 7.0-99. Informatsionno-bibliotechnaya deyatel'nost', bibliografiya. Terminy i opredeleniya [Information and library activities, bibliography. Terms and Definitions]. — Data vved. 2000-07-01 (in Russian).

Гуларян А. Б. Принцип «избыточности» как основа построения семантических систем // Грани познания. — 2010. — № 1. — С. 1–3.

Gularyan A. B. Printsip «izbytochnosti» kak osnova postroeniya semanticheskikh sistem [The principle of «redundancy» as the basis for the construction of semantic systems] // Grani poznaniya. — 2010. — No. 1. — P. 1–3 (in Russian).

Давыдова А. В. Новые технологии во внутреннем аудите // Деньги и кредит. — 2017. — № 2. — С. 43–44.

Davydova A. V. Novye tekhnologii vo vnutrennem audite [New technologies in internal audit] // Den'gi i kredit. — 2017. — No. 2. — P. 43–44 (in Russian).

Дьяконов А. Г. Анализ данных, обучение по прецедентам, логические игры, системы WEKA, RapidMiner и MatLab: практикум на ЭВМ кафедры математических методов прогнозирования. — М.: Изд. отдел ф-та ВМК МГУ им. М. В. Ломоносова, 2010. — 278 с.

D'yakonov A. G. Analiz dannykh, obuchenie po pretsedentam, logicheskie igry, sistemy WEKA, RapidMiner i MatLab: praktikum na EVM kafedry matematicheskikh metodov prognozirovaniya [Data analysis, teaching by precedents, logic games, WEKA, RapidMiner and MatLab systems: computer workshop of the Department of Mathematical Forecasting Methods]. — M.: Izd. otdel f-ta VMK MGU im. M. V. Lomonosova, 2010. — 278 p. (in Russian).

Копчёнова Л. А., Орлова В. Е., Селезнёва С. В. Информационный взрыв и его влияние на современного человека // Научные труды SWorld. — 2013. — Т. 44, № 1. — С. 10–12.
Kopchenova L. A., Orlova V. E., Selezneva S. V. Informatsionnyi vzryv i ego vliyanie na sovremennogo cheloveka [Information explosion and its impact on modern human] // Nauchnye trudy SWorld. — 2013. — T. 44, no. 1. — P. 10–12 (in Russian).

Маннинг К. Д., Рагхаван П., Шютце Х. Введение в информационный поиск. — М.: ООО «Вильямс», 2011. — 528 с.
Manning C. D., Schütze H., Raghavan P. Introduction to information retrieval. — Cambridge: Cambridge University Press, 2008. — Vol. 39. — P. 234–265. (Russ. ed.: *Manning K. D., Ragkhavan P., Shyuttse Kh.* Vvedenie v informatsionnyi poisk. — M.: OOO «Vil'yams», 2011. — 528 p.)

Мусаев А. А., Григорьев Д. А. Формализованная постановка и краткий обзор технологий извлечения знаний из текстовых документов в задачах управления финансовыми активами // II Всероссийская научно-практическая конференция с международным участием «Техника и технология современных производств», Пензенский государственный университет, 2021.
Musaev A. A., Grigoriev D. A. Formalizovannaya postanovka i kratkiy obzor tehnologiy izvlecheniya znaniy iztekstovih documentov v zadachah upravleniya finansovymi aktivami [A formalized formulation and a brief overview of the technologies extracted learning knowledge from text documents in the tasks of managing financial as sets] // II All-Russian scientific and practical conference with international participation «Technique and technology of modern production», Penza State University, 2021 (in Russian).

Наумов В. Н. Анализ данных и машинное обучение. Методы и инструментальные средства. — СПб.: ИПЦ СЗИУ РАНХиГС, 2020. — 260 с.
Naumov V. N. Analiz dannykh i mashinnoe obuchenie. Metody i instrumental'nye sredstva [Data analysis and machine learning. Methods and tools]. — SPb.: IPTs SZIU RANKhiGS, 2020. — 260 p. (in Russian).

Тимофеев А. Г., Лебединская О. Г. Data mining и big data в бизнес-аналитике цифровой трансформации государственного и корпоративного управления // Управление экономическими системами: электронный научный журнал. — 2017. — № 9 (103). — URL: http://uecs.ru/uecs-103-1032017/item/4533-data-mining-big-dat- (дата обращения: 14.04.2021).
Timofeev A. G., Lebedinskaya O. G. Data mining i big data v biznes-analitike tsifrovoi transformatsii gosudarstvennogo i korporativnogo upravleniya [Business analysis under the conditions of digital transformation of state and corporate governance] // Management of economic systems: scientific economic journal. — 2017. — No. 9 (103). — Available at: http://uecs.ru/uecs-103-1032017/item/4533-data-mining-big-dat- (accessed: 14.04.2021) (in Russian).

Фомичева Т. Л., Магомедов Р. М., Викулина Е. А. Применение методов интеллектуального анализа данных и machine learning в борьбе с мошенничеством в банках // Самоуправление. — 2019. — Т. 2, № 3. — С. 337–339.
Fomicheva T. L., Magomedov R. M., Vikulina E. A. Primenenie metodov intellektual'nogo analiza dannykh i machine learning v bor'be s moshennichestvom v bankakh [Application of Data Mining and Machine Learning Methods to Bank Fraud Investigations] // Samoupravlenie. — 2019. — Vol. 2, no. 3. — P. 337–339 (in Russian).

Agarwal M. An Overview of Natural Language Processing // International Journal for Research in Applied Science and Engineering Technology. — 2019. — No. 7. — P. 2811–2813.

Ahonen H. et al. Applying data mining techniques for descriptive phrase extraction in digital document collections // Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries-ADL'98. — 1998. — No. 2-11.

Allahyari M. et al. A brief survey of text mining: Classification, clustering and extraction techniques // arXiv preprint arXiv:1707.02919. — 2017.

Aloysius G., Binu D. An approach to products placement in supermarkets using PrefixSpan algorithm // Journal of King Saud University-Computer and Information Sciences. — 2013. — Vol. 25, no. 1. — P. 77–87.

Anshika S., Udayan G. Text Mining: A Burgeoning technology for knowledge extraction // International Journal of Scientific Research Engineering & Technology (IJSRET). — 2013. — Vol. 1, no. 12. — P. 022–026.

Araci D. Finbert: Financial sentiment analysis with pre-trained language models // arXiv preprint arXiv:1908.10063. — 2019.

Atika M., Ali A., Ahmer S. Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and Categorization // International Journal of Multimedia and Ubiquitous Engineering. — 2009. — Vol. 4, no. 2.

Bartschat A., Reischl M., Mikut R. Data mining tools // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. — 2019. — Vol. 9, no. 4. — P. e1309.

Bakshi R. K. et al. Opinion mining and sentiment analysis // 2016 3rd international conference on computing for sustainable global development (INDIACom). — IEEE, 2016. — P. 452–455.

Berkhin P. Survey of clustering data mining techniques // Technical Report, Accrue Software. Inc. TR, San Jose, USA. — 2002.

Bollen J., Mao H., Zeng X. Twitter mood predicts the stock market // Journal of computational science. — 2011. — Vol. 2, no. 1. — P. 1–8.

Borgelt C. Frequent item set mining // Wiley interdisciplinary reviews: data mining and knowledge discovery. — 2012. — Vol. 2, no. 6. — P. 437–456.

Brown T. B. et al. Language models are few-shot learners // arXiv preprint arXiv:2005.14165. — 2020.

Buzmakov A., Kuznetsov S. O., Napoli A. Fast generation of best interval patterns for nonmonotonic constraints // Joint European Conference on Machine Learning and Knowledge Discovery in Databases. — Springer, Cham, 2015. — P. 157–172.

Calvillo E. A. et al. Searching research papers using clustering and text mining // CONIELECOMP 2013, 23rd International Conference on Electronics, Communications and Computing. — IEEE, 2013. — P. 78–81.

Chambers N. et al. Learning alignments and leveraging natural logic / In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. — 2007. — P. 165–170.

Chantrapornchai C., Tunsakul A. Information Extraction Tasks based on BERT and SpaCy on Tourism Domain // ECTI Transactions on Computer and Information Technology (ECTI-CIT). — 2021. — Vol. 15, no. 1. — P. 108–122.

Chen D. et al. Group, extract and aggregate: Summarizing a large amount of finance news for forex movement prediction // arXiv preprint arXiv:1910.05032. — 2019.

Chowdhury G. G. Natural language processing // Annual review of information science and technology. — 2003. — Vol. 37, no. 1. — P. 51–89.

Collobert R., Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning // Proceedings of the 25th international conference on Machine learning. — 2008. — P. 160–167.

De la Torre C. J. et al. Text mining: intermediate forms on knowledge representation // EUSFLAT Conf. — 2005. — P. 1082–1087.

Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding // arXiv preprint arXiv:1810.04805. — 2018.

Ding X. et al. Knowledge-driven event embedding for stock prediction // Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers. — 2016. — P. 2133–2142.

Fradkin D., Moerchen F. Margin-closed frequent sequential pattern mining // Proceedings of the ACM SIGKDD Workshop on Useful Patterns. — 2010. — P. 45–54.

Gantz J., Reinsel D. Extracting value from chaos // IDCiview. — 2011. — Vol. 1142, no. 2011. — P. 1–12.

Ghafari S. M., Tjortjis C. A survey on association rules mining using heuristics // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. — 2019. — Vol. 9, no. 4. — P. e1307.

Ghorbani A. et al. Towards automatic concept-based explanations // arXiv preprint arXiv:1902.03129. — 2019.

Ghosh S., Roy S., Bandyopadhyay S. K. A tutorial review on Text Mining Algorithms // International Journal of Advanced Research in Computer and Communication Engineering. — 2012. — Vol. 1, no. 4. — P. 7.

Gupta V. et al. A survey of text mining techniques and applications // Journal of emerging technologies in web intelligence. — 2009. — Vol. 1, no. 1. — P. 60–76.

Hipp J., Güntzer U., Nakhaeizadeh G. Algorithms for association rule mining — a general survey and comparison // ACM sigkdd explorations newsletter. — 2000. — Vol. 2, no. 1. — P. 58–64.

Hosseininasab A., van Hoeve W.J., Cire A.A. Constraint-based sequential pattern mining with decision diagrams // Proceedings of the AAAI Conference on Artificial Intelligence. — 2019. — Vol. 33, no. 01. — P. 1495–1502.

ISO/IEC 2382:2015 Information technology. — Vocabulary: knowledge concerning objects, such as facts, events, things, processes, or ideas, including concepts, that within a certain context has a particular meaning.

Jacobs G., Hoste V. Extracting Fine-Grained Economic Events from Business News // Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation. — 2020. — P. 235–245.

Janani R., Vijayarani S. Text mining research: A survey // Int. J. Innov. Res. Comput. Commun. Eng. — 2016. — Vol. 4, no. 4. — P. 6564–6571.

Jiang L. et al. Target-dependent twitter sentiment classification // Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. — 2011. — P. 151–160.

Jurafsky D., Martin J. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 3d edition draft, 2020. — Available at: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf (accessed: 14.04.2021).

Kalyan K.S., Sangeetha S. Bertmcn: Mapping colloquial phrases to standard medical concepts using bert and highway network // Artificial Intelligence in Medicine. — 2021. — Vol. 112. — P. 102008.

Kaytoue M., Kuznetsov S.O., Napoli A. Revisiting numerical pattern mining with formal concept analysis // arXiv preprint arXiv:1111.5689. — 2011.

Kitsuregawa M., Nishida T. Special issue on information explosion // New Generation Computing. — 2010. — Vol. 28, no. 3. — P. 207–215.

Khurana D. et al. Natural language processing: State of the art, current trends and challenges // arXiv preprint arXiv:1708.05148. — 2017.

Kumar L., Bhatia P.K. Text mining: concepts, process and applications // Journal of Global Research in Computer Science. — 2013. — Vol. 4, no. 3. — P. 36–39.

Lee G. Concept-based method for extracting valid subsets from an EXPRESS schema // Journal of Computing in Civil Engineering. — 2009. — Vol. 23, no. 2. — P. 128–135.

Liu B. et al. Sentiment analysis and subjectivity // Handbook of natural language processing. — 2010. — Vol. 2, no. 2010. — P. 627–666.

Liu Z. et al. Finbert: A pre-trained financial language representation model for financial text mining // Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI. — 2020. — P. 5–10.

Lodhi H. et al. Text classification using string kernels // Journal of Machine Learning Research. — 2002. — Vol. 2, no. Feb. — P. 419–444.

Minaee S. et al. Deep learning based text classification: A comprehensive review // arXiv preprint arXiv:2004.03705. — 2020.

Nasa D. Text mining techniques — A survey // International Journal of Advanced Research in Computer Science and Software Engineering. — 2012. — Vol. 2, no. 4. — P. 50–54.

Nalini K., Sheela L.J. Survey on text classification // International Journal of Innovative Research in Advanced Engineering. — 2014. — Vol. 1, no. 6. — P. 412–417.

Prokasheva O., Onishchenko A., Gurov S. Classification methods based on formal concept analysis // FCAIR 2012 – Formal Concept Analysis Meets Information Retrieval. — 2013. — P. 95.

Radford A. et al. Language models are unsupervised multitask learners // OpenAI blog. — 2019.— Vol. 1, no. 8. — P. 9.

Rajak A., Gupta M. K. Association rule mining: applications in various areas // Proceedings of international conference on data management, Ghaziabad, India. — 2008. — P. 3–7.

Ratra R., Gulia P. Experimental Evaluation of Open Source Data Mining Tools (WEKA and Orange) // International Journal of Engineering Trends and Technology (IJETT). — 2021. — Vol. 68, no. 8.

Sagayam R., Srinivasan S., Roshni S. A survey of text mining: Retrieval, extraction and indexing techniques // International Journal of Computational Engineering Research. — 2012. — Vol. 2, no. 5. — P. 1443–1446.

Salton G., Buckley C. Term-weighting approaches in automatic text retrieval // Information processing & management. — 1988. — Vol. 24, no. 5. — P. 513–523.

Sert O. C. et al. Analysis and prediction in sparse and high dimensional text data: The case of Dow Jones stock market // Physica A: Statistical Mechanics and its Applications. — 2020. — Vol. 545. — P. 123752.

Sesen M. B., Romahi Y., Li V. Natural language processing of financial news // Big Data and Machine Learning in Quantitative Investment. — 2019. — P. 185.

Stavrianou A., Andritsos P., Nicoloyannis N. Overview and semantic issues of text mining //ACM Sigmod Record. — 2007. — Vol. 36, no. 3. — P. 23–34.

Steinbach M., Karypis G., Kumar V. A Comparison of Document Clustering Techniques, Computer Science & Engineering (CS&E) Technical Reports, University of Minnesota, 2000.

Sun A., Lachanski M., Fabozzi F. J. Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction // International Review of Financial Analysis. — 2016. — Vol. 48. — P. 272–281.

Tax N. et al. Mining local process models // Journal of Innovation in Digital Ecosystems. — 2016. — Vol. 3, no. 2. — P. 183–196.

Thelwall M. et al. Sentiment strength detection in short informal text // Journal of the American society for information science and technology. — 2010. — Vol. 61, no. 12. — P. 2544–2558.

Thilagavathi K., Shanmuga V. A survey on text mining techniques // Int. J. Adv. Res. Comput. Sci. Robot. — 2014. — Vol. 2, no. 10. — P. 41–50.

Thirumuruganathan S. et al. Beyond itemsets: mining frequent featuresets over structured items // Proceedings of the VLDB Endowment. — 2014. — Vol. 8, no. 3. — P. 257–268.

Umajancy S., Thanamani A. S. An analysis on text mining-text retrieval and text extraction // International Journal of Advanced Research in Computer and Communication Engineering. — 2013. — Vol. 2, no. 8.

Wu S. T. et al. Automatic pattern-taxonomy extraction for web mining // IEEE/WIC/ACM International Conference on Web Intelligence (WI'04). — IEEE, 2004. — P. 242–248.

Xing F., Hoang D. H., Vo D. V. High-frequency news sentiment and its application to forex market prediction // Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS). — 2020.

Xing F. Z., Cambria E., Welsch R. E. Natural language based financial forecasting: a survey // Artificial Intelligence Review. — 2018. — Vol. 50, no. 1. — P. 49–73.

Young T. et al. Recent trends in deep learning based natural language processing // IEEE Computational Intelligence Magazine. — 2018. — Vol. 13, no. 3. — P. 55–75.

Zeng C., Naughton J. F., Cai J. Y. On differentially private frequent itemset mining // The VLDB journal: very large data bases: a publication of the VLDB Endowment. — 2012. — Vol. 6, no. 1. — P. 25.

Zhong Z. et al. SemRegex: A semantics-based approach for generating regular expressions from natural language specifications / Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. — 2018.

Приложение. Обобщенные характеристики инструментов извлечения знаний

Название	Разработчик	Язык	Дата релиза	Лицензия	Тип ПО	Область применения
WEKA	University of Waikato (Новая Зеландия)	Java	1993	GNU GPLv3	DMS, LIB	IE, ML
Orange	University of Ljubljana (Швейцария)	C++, Python	1996	GNU GPLv3	DMS	IE, ML
GATE	University of Sheffield (Англия)	Java	1996	GNU GPLv3	SOL	IE, TM
RapidMiner	University of Dortmund (Германия)	Java, R, Python, WEKA	2001	AGPL-3.0 / Commercial	DMS	IE, ML, деревья решений (DT)
NLTK	Team NLTK	Python	2001	Apache License v2	LIB	IE, TM, NLP
UIMA	IBM, Apache	Java, C++	2002	Apache License v2	SOL	IE, TM
OpenNLP	Apache Software Foundation	Java	2004	Apache License v2	LIB	IE, TM, NLP
KNIME	University of Konstanz (Германия)	Java, R, WEKA	2006	GNU GPLv3 / Commercial	DMS	ML, IE, DT, правила ассоциаций
Gensim	RARE Technologies Ltd.	Python	2008	GNU LGPLv2.1	LIB	IE, TM, IR
Stanford CoreNLP	Stanford University (США)	Java	2010	GNU GPLv3	LIB	IE, TM, NLP
SpaCy	Explosion AI	Python	2015	MIT License	LIB	IE, TM, NLP, трансформеры
Томита-парсер	«Яндекс»	DSL	2012	MPL 2.0	BIN	IE, TM, NLP
Natasha	Лаборатория анализа данных А. Кукушкина	Python	2016	MIT License	LIB	IE, TM, NLP
ABBYY FlexiCapture	ABBYY	C, C++, Java	2012	Commercial	SOL	IE, NLP, ML
RCO Fact Extractor	RCO	C++	2017	Commercial	LIB	IE, NLP, ML
IBM Watson NLP/NLU	IBM	Web API	2015	Commercial	DMS, LIB	IE, NLP, NLU, ML
Google Cloud Natural Language	Google	Web API	2016	Commercial	DMS, LIB	IE, NLP, NLU, ML
Microsoft Text Analytics	Microsoft	Web API	2016	Commercial	DMS, LIB	IE, NLP, NLU, ML
Deep Pavlov	МФТИ (Москва)	Python	2017	Apache License v3	SOL, LIB	IE, NLP, NLU, трансформеры, DL
NLP Architect	Intel	Python	2018	Apache License v2	SOL, LIB	IE, NLP, NLU, DL