УДК 004.912
DOI 10.17513/snt.40387

ПОИСК И ИДЕНТИФИКАЦИЯ ТЕКСТОВ ОПРЕДЕЛЕННОЙ СЕМАНТИЧЕСКОЙ НАПРАВЛЕННОСТИ В ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ПОТОКАХ

Вишняков Ю.М., Вишняков Р.Ю.

ФГАОУ ВО «Кубанский государственный университет», Краснодар, e-mail: jury.vishnyakov@gmail.com

В обработке естественно-языковой информации актуальна проблема выявления текстов определенной семантической направленности и определения их источников. Это требуется в анализе новостных потоков, чатов мессенджеров, социальных сетей, проверке документов на плагиат и других подобных задачах. Целью работы является обоснование концептуальной модели выявления в естественно-языковых потоках текстов определенной семантической направленности по формальным описаниям их источников. Анализ известных подходов показал потребность в собственном инструментарии для решения проблемы. В работе предлагается семантическую направленность задавать сценариями языка формальной грамматики гипотетического семантического объекта, сценарии представлять последовательностями характеристик семантического словаря и направленность текста определять семантической близостью сценарию. Бесконечность языка сценариев и отсутствие информации об исходном тексте исключают простой перебор, поэтому предполагаемый сценарий конструируется. Процесс организуется последовательным определением семантического сходства токенов текста характеристикам и их сборкой в предполагаемый сценарий, проверяемый на принадлежность языку. Для семантического сравнения текстов и сценариев сконструированы функции семантического подобия, общий и частный алгоритмы выявления текстов определенной семантической направленности. В общем алгоритме разбор сводится к построению вывода в формальной грамматике, для регулярных грамматик разбор выполняется системой переходов. Для ускорения сборка предполагаемого сценария совмещается с грамматическим разбором и используется механизм бек-трекинга. Точность алгоритмов определяется фактической близостью текстов сценариям. В работе приводится состав разработанного программного комплекса, тестирование которого подтверждает теоретические результаты. Исследование развивает фундаментальные основы математического моделирования естественно-языковой обработки и предлагает новые эффективные вычислительные алгоритмы для комплексов проблемно-ориентированных программ.

Ключевые слова: естественно-языковая обработка, текстовый поток, семантика, формальная модель, формальная грамматика, формальный язык, вывод, система переходов, алгоритм

Исследование выполнено при финансовой поддержке Кубанского научного фонда в рамках научно-инновационного проекта «НИП-20.1.4».

SEARCH AND IDENTIFICATION OF TEXTS WITH A SPECIFIC SEMANTIC ORIENTATION IN NATURAL LANGUAGE STREAMS

Vishnyakov Yu.M., Vishnyakov R.Yu.

Kuban State University, Krasnodar, e-mail: jury.vishnyakov@gmail.com

In natural language processing, an important challenge is the identification of texts with a specific semantic orientation and the determination of their sources. This is essential for tasks such as analyzing news streams, messenger chats, and social media, as well as for plagiarism detection and other similar applications. The objective of this study is to substantiation of a conceptual model for identifying texts of a certain semantic orientation in natural language flows based on formal descriptions of their sources. A review of existing approaches revealed the necessity of creating a dedicated toolkit to address this problem. The proposed approach defines semantic orientation through scenarios of a formal grammar language representing a hypothetical semantic object. These scenarios are expressed as sequences of characteristics from a semantic dictionary, and the orientation of a text is determined by its semantic proximity to a given scenario. Due to the infinite nature of the scenario language and the lack of prior information about the original text, exhaustive search is impractical. Instead, a hypothesized scenario is constructed dynamically. The process involves sequentially determining the semantic similarity of text tokens to characteristics and assembling them into a hypothesized scenario, which is then verified for conformity to the language. To compare the semantics of texts and scenarios, semantic similarity functions have been developed, along with general and specialized algorithms for detecting texts with a specific semantic orientation. The general algorithm frames parsing as the construction of a derivation in a formal grammar, while for regular grammars, parsing is performed using a transition system. To enhance efficiency, the assembly of the hypothesized scenario is integrated with grammatical parsing, employing a backtracking mechanism. The accuracy of the algorithms is determined by the actual proximity of texts to the given scenarios. The study also presents the structure of the developed software system, whose testing confirms the theoretical findings. This research contributes to the fundamental principles of mathematical modeling in natural language processing and introduces new, efficient computational algorithms for problem-oriented software solutions.

Keywords: natural language processing, text stream, semantics, formal model, formal grammar, formal language, derivation, transition system, algorithm

The study was carried out with the financial support of the Kuban Science Foundation within the framework of the scientific and innovative project «NIP-20.1.4».

Введение

В настоящее время в обработке естественно-языковой информации больших объемов возникла потребность выявления текстов определенной семантической направленности и определения их источников, что зачастую необходимо в анализе новостных потоков, чатов мессенджеров, социальных сетей, проверке документов на плагиат, установлении авторства и других подобных задачах. Однако, несмотря на наличие современного инструментария [1, с. 79–181] и его специальные приложения [2, с. 130–190], позволяющие достаточно быстро и эффективно разрабатывать программы, проблема нуждается в хорошей теоретической проработке и создании собственного формального подхода для ее решения.

Анализ показывает, что в области естественно-языковой обработки существует большое число различных исследований, которые находят отражение во многих публикациях. Если попытаться условно классифицировать их по направлениям, то, наверное, в первую группу следует включить работы фундаментального характера по всему спектру естественно-языковой обработки, например [3, с. 60–93]. Вторую группу можно составить из работ, направленных на поиск универсальных решений на основе нейросетевых технологий [4]. В третью группу сводятся работы по методам извлечения информации из текстов [5; 6]. Четвертую группу можно образовать по семантической кластеризации областей знаний и их классификации [7]. Группа исследований в нетрадиционной постановке также многочисленна, например обработка естественно-языковых запросов для точных областей знаний [8]. В нее можно добавить области обучения [9; 10] и мультимедиа [11]. Отдельную группу образуют работы по семантическим моделям конкретных частных задач [12–14]. Безусловно, вышеперечисленным не исчерпывается все многообразие исследований, однако анализ показывает, что все они только частично в той или иной мере касаются проблемы поиска естественно-языковых потоков текстов определенной семантической направленности, само решение проблемы видится в создании собственных теоретические наработок.

Следует также обратить внимание на точность естественно-языковой обработки. Она зависит от множества факторов, ключевые из которых представляют используемая модель семантики и затраты на ее реализацию. На сегодня наиболее распространена и изучена частотная модель семантики, ее суть выражает сентенция – текст тем релевантнее запросу, чем чаще его слова входят в этот текст. Типичные частотные методы – «мешок слов», алгоритмы TF-IDF, векторные модели, нейросетевые и пр. Однако частотная модель заранее должна быть настроена на статистические характеристики обрабатываемых текстов. Процесс требует большого объема тщательно составленных обучающих выборок, но и после этого доверие к точности результатов обработки может быть подвергнуто сомнению. В контексте повышения точности обработки естественно-языковых данных на многочисленных научных форумах и в публикациях уже давно дискутируется вопрос о переходе на иные модели семантики, поскольку затраты на улучшение частотной модели зачастую не приводят к ожидаемым результатам. В этом ключе авторы развивают вычислительную модель представления семантики [15], положенную в основу предлагаемого исследования. Непосредственно сама предлагаемая работа направлена на решение проблемы поиска и выявления текстов определенной семантической направленности в естественно-языковых потоках.

Цель исследования – обоснование концептуальной модели выявления в естественно-языковых потоках текстов определенной семантической направленности по формальным описаниям их источников.

Материалы и методы исследования

Трудности решения проблемы состоят в том, что понятие «определенная семантическая направленность» имеет вербальную формулировку и соотнесение с ней смыслового содержания текста понятно в человеческом контексте. Однако формальное решение предполагает отображение проблемы в точные формальные модели, процедуры и алгоритмы, которые могут быть программно реализованы. Для этого в исследовании предлагается использовать вычислительную теорию семантической интерпретации и теорию формальных языков и грамматик. Обсудим прежде несколько формальных понятий.

Под естественно-языковым потоком будем понимать последовательность неделимых по смыслу текстовых элементов (токенов – предложений или их фрагментов) вида $T = t_1 t_2 \ldots t_m$, где t_i – токен.

Семантическим объектом будем считать некий гипотетический объект, который, исходя из своих внутренних целей и задач, из токенов множества $P = \{p_1, p_2, \ldots, p_n\}$ по определенным правилам может сконструировать осмысленный текст α_i из множества текстовых последовательностей $L(P) = \{\alpha_1, \alpha_2, \ldots \alpha_m\}$, $L(P) \subset P^*$

и $P^* = P^0 \cup P^1 \cup \ldots \cup P^l \cup \ldots$, где $P^0 = \{\lambda\}$ – аксиома и λ – пустая цепочка. Совместно множество P и правила (способы) конструирования последовательностей $L(P)$ назовем семантическим описанием объекта.

Для семантического объекта множество P играет роль смыслового лексикона, из элементов которого конструируются текстовые последовательности (тексты) определенного смыслового содержания. Элементы $p_i \in P$ назовем характеристиками семантического объекта, само P – характеристическим множеством и сконструированный текст $\alpha_i \in L(P)$ – сценарием [6].

С учетом данных понятий формальная постановка задачи выглядит следующим образом. Пусть имеется некоторый текстовый поток T и заданный характеристическим множеством P и множеством сценариев $L(P)$ семантический объект. Требуется проверить присутствие в текстовом потоке текста, семантически подобного сценарию $\alpha_i \in L(P)$, и вычислить степень такого подобия.

1. *Семантическое подобие текстовых элементов*

Обсудим семантическое подобие двух текстовых элементов (токенов) в вычислительной модели семантики [15]. Будем считать, что и токены из T, и элементы из P целостны по смыслу и представляют предложения или их фрагменты. Например, пусть имеются два текстовых элемента *p = «международное признание образовательных программ российских вузов» и t = «образовательные программы российских вузов нуждаются в международном признании»*. Обозначим множества их смысловых значений через $S(p)$ и $S(t)$ *соответственно* и для наглядности сопоставим в порядке следования в тексте элементе p словам малые буквы латинского алфавита:

$$p = "\underbrace{\text{международное}}_{a}\ \underbrace{\text{признание}}_{b}\ \underbrace{\text{образовательных}}_{c}\ \underbrace{\text{программ}}_{d}\ \underbrace{\text{российских}}_{e}\ \underbrace{\text{вузов}}_{f}" . \quad (1)$$

Тогда формульное представление функционала смысла $S(p)$ в нотации обратной польской записи [15] имеет вид

$$|E| = \left\{ \propto = p_{r1} p_{r2} \ldots p_{rk}, \text{ где } p_{ri} \in P u \propto \in L(P) \right\}, \quad (2)$$

где $\overleftarrow{\cap}$ – операция контекстного уточнения смысла, а k – арность операции.

Вычислительную процедуру функционала смысла $S(p)$ представляет семантическая схема (рис. 1).

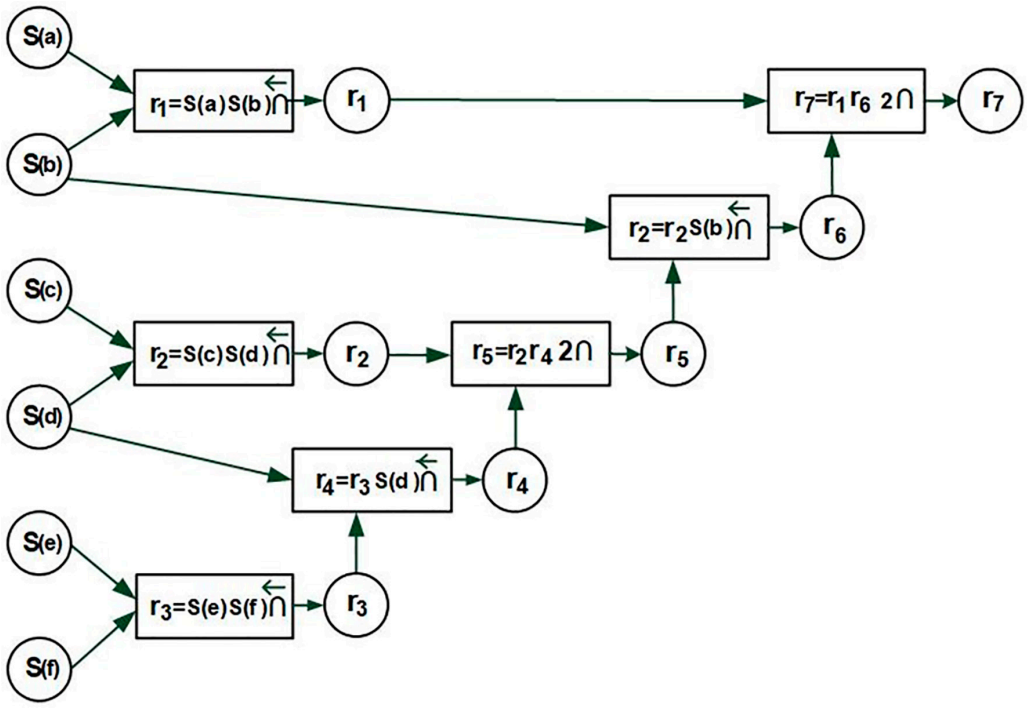

Рис. 1. Семантическая схема текстового элемента p
Источник: составлено авторами по [15]

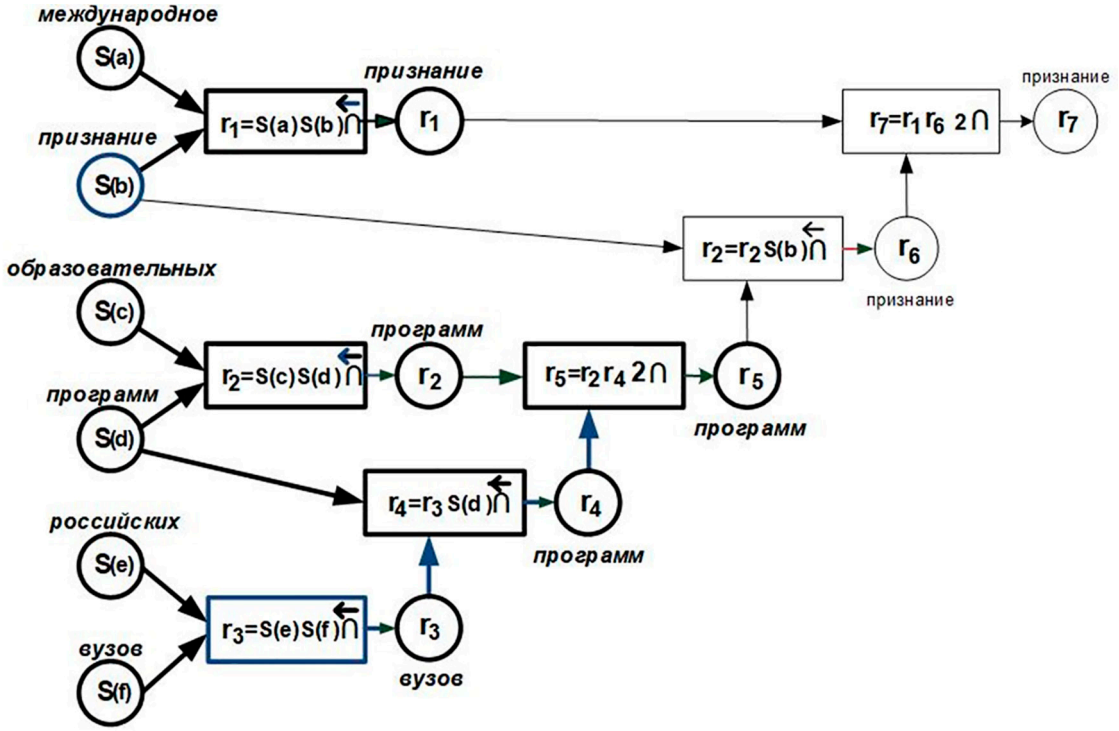

Рис. 2. Семантическая схема текстового элемента t
Источник: составлено авторами по [15]

На семантической схеме самые левые круглые вершины соответствуют ее входам, прямоугольные (элементы смысла) – операциям, результаты вписаны в круглые вершины на выходах элементов смысла. Подобное формульное представление смыслового функционала можно сконструировать для текстового элемента *t*, оно показано на рис. 2.

В качестве меры подобия (сходства, близости) семантических значений $S(p)$ и $S(t)$ выбирается критерий вида

$$0 \le Cprox(S(p),S(t)) \le 1.$$

Если $Cprox = 0$, то семантическая близость между *p* и *t* отсутствует, при $Cprox = 1$ имеет место полное семантическое сходство, в иных случаях семантическая близость частичная. В силу несимметричности операции контекстного уточнения смысла имеет место утверждение

$$Cprox(S(p),S(t)) \ne Cprox(S(t),S(p)).$$

Непосредственно критерий *Cprox* представляется выражением

$$Cprox(S(p),S(t)) = q / m, \qquad (3)$$

где *q* – число элементов смысла семантической схемы элемента *p*, совпадающих с элементами смысла семантической схемы элемента *t,* и *m* – общее число элементов смысла семантической схемы *p.* В рассматриваемом случае семантическая схема *p* содержит 6 слов и 7 элементов смысла (прямоугольных вершин), то есть m = 7.

На семантической схеме элемента *t* (рис. 2) жирными линиями выделены элементы смысла, совпадающие с элементами смысла семантической схемы элемента *p* (рис. 1), и число таких совпадений $q = 5$.

Таким образом, значение семантического подобия текстовых элементов *p* и *t* $Cprox(S(p),S(t)) = 5/7 = 0{,}71$ и достаточно высоко. Если близость определять на основе частотной модели, то она равнялась бы единице, поскольку все слова элемента *p* входят в элемент *t*. В действительности утверждение о полной семантической близости является не в полной мере истинным.

2. Семантическое подобие текстовых последовательностей, функция подобия

Пусть текстовый поток представлен последовательностью токенов $T = t_1 t_2 \ldots t_m$. Обозначим множество токенов, входящих в последовательность *T*, через $T' = \{t_1, t_2, \ldots, t_m\}$, и построим отображение $Q = P \rightarrow T'$ таким образом, чтобы элементами отображения были пары (p_i, t_j), где $p_i \in P$, $t_j \in T'$. Для каждой пары определим меру близости

как $Cprox(p_i,t_j) = w$ и дополним ею пару в отображении. После этого модифицированное отображение принимает вид

$$Q=\left\{\left(p_i,t_j,w_k\right), p_i \in P,\ t_j \in T', w_k = Cprox\left(p_i,t_j\right), 0 < w_k\right\} \quad (4)$$

Придадим мере близости w_k следующую трактовку, будем считать ее мерой присутствия характеристики p_i в текстовом потоке, а токен $t_j = (p_i,w_k)$, являющийся ее образом в отображении, будем считать семантическим следом характеристики p_i в текстовом потоке.

Пусть в языке сценариев $L(P)$ семантического объекта имеется некоторый сценарий $\alpha = p_{i1}p_{i2}\dots p_{ik}$, $\alpha \in P$ и существует его отображение в текстовый поток вида $T = \dots t_{r1}\dots t_{r2}\dots t_{rk}$ …. Если из потока T удалить незначащие токены, обладающие нулевой степенью подобия с характеристиками, то получится некоторая последовательность токенов $\alpha' = t_{r1}t_{r2}\dots t_{rk}$, которая представляет собой отображение сценария α на текстовый поток, будем ее считать семантическим следом данного сценария α в текстовом потоке. В дальнейшем подобное удаление незначащих токенов из текстового потока назовем его нормализацией.

Для оценки семантического подобия сценария α и семантического следа α' сконструируем функцию подобия $F\left(\alpha,\propto'\right)$. Значения функции нормализуем на интервале $0 \le F\left(\alpha,\propto'\right) \le 1$, а ее аргументами будем считать значения критериев близости характеристик и токенов, входящих в последовательности α и α'. С учетом данного обстоятельства функция подобия принимает вид

$$F\left(\alpha,\propto'\right) = F(\left(Cprox\left(q_{i1},t_{r1}\right),Cprox\left(q_{i2},t_{r2}\right),\dots,Cprox\left(q_{ik},t_{rk}\right)\right)=F\left(w_{r1},w_{r2},\dots w_{rk}\right) \quad (5)$$

Исходя из практических целей, для F можно выбирать различные виды аппроксимаций, например учитывать значимость характеристик, их роли и места в сценарии, обеспечивая тем самым тонкую настройку функции на различные ситуации. Для общего понимания можно F принять как средневзвешенную величину вида

$$F\left(\alpha,\propto'\right)=\frac{1}{k}\sum_{i=1}^{k} w_{ri}\,. \quad (6)$$

Как показывает практика, этого вполне достаточно во многих ситуациях. Очевидно, что значение функции подобия $F\left(\alpha,\propto'\right)$ также можно считать мерой присутствия сценария α семантического объекта в текстовом потоке.

3. Поиск семантических следов в текстовом потоке

Пусть своим характеристическим множеством P и множеством сценариев $L(P)$ задан семантический объект и имеется некоторый текстовый поток T. Требуется установить, присутствует ли сценарий α семантического объекта в текстовом потоке T, и оценить меру такого присутствия.

Подход к решению сводится к подбору из множества $L(P)$ такого сценария α, семантическим следом которого в потоке T является последовательность токенов α' с максимальной функцией подобия $F\left(\alpha,\propto'\right)$. Трудность состоит в том, что ни сценарий α, ни его семантический след α' заранее не известны, их требуется найти.

При конечном множестве сценариев $L(P)$ решение задачи можно свести к тривиальному методу перебора, но случай бесконечного множества требует иных подходов. Первый из них лежит на поверхности и представляет собой нейросетевой подход. Однако с ним не все просто – качественное обучение нуждается в тщательно подобранных статистических данных большого объема, которых зачастую нет, само обучение также нуждается в немалом времени, и, самое главное, никакой уверенности в точности результатов на реальных входных данных нет.

Более оптимистичен и точен формально-грамматический поход. В нем множество сценариев $L(P)$ рассматривается как язык некоей формальной грамматики $G[Z]$, в которой Z – начальный символ, множество P – терминальный словарь и сценарий α – предложение. Тогда процесс распознавания сценария α_i сводится к построению вывода $Z \Rightarrow +\alpha_i$ из начального символа грамматики по типу синтаксического анализа с одновременным вычислением функции подобия. Здесь под выводом $\Rightarrow +$ понимается транзитивное замыкание отношения простого вывода на множестве цепочек грамматики, но для понимания вполне достаточно под ним понимать возможность вывода сценария α_i из начального символа грамматики.

Сконструируем для данного подхода укрупненный алгоритм выявления семантического следа объекта. Пусть текстовый поток нормализован и представлен по-

следовательностью токенов $T = t_1t_2\ldots t_l$ и $P = \{p_1, p_2,\ldots, p_n\}$ – характеристическое множество семантического объекта, а $L(P)$ – множество его сценариев, являющееся языком грамматики $G[Z]$. Для реализации алгоритма введем вспомогательную текстовую переменную *alpha*, в которой будем собирать искомый сценарий семантического объекта α, начальным значением переменной положим пустую цепочку λ. Если характеристика p_i использовалась при составлении сценария (катенирована с переменной $alpha+p_i$), то она помечается меткой использования. Также введем вспомогательную переменную *s,* в которой будем накапливать значение близостей отдельных следов характеристик из *alpha* и ее начальным значением положим ноль.

Алгоритм:

1. Установить начальные значения

$$alpha = \lambda \text{ и } s = 0.$$

2. Последовательно для каждого токена t_i, $i = 1, 2, \ldots l$ текстового потока $T = t_1t_2\ldots t_l$:

2.1. Снять метки использования с характеристик множества $P = \{p_1, p_2,\ldots, p_n\}$.

2.2. Выбрать из непомеченных характеристик множества P характеристику с максимальным ненулевым значением $w_j = Cprox(p_j, t_i)$.

2.3. Выполнить катенацию переменной *alpha* с характеристикой p_j вида $alpha: = alpha + p_j$, пометить характеристику p_j как использованную.

2.4. Если вывод $Z\Rightarrow+alpha$ существует, то переменную s увеличить $s: = s+w_j$, иначе характеристику откатенировать $alpha: = alpha - p_j$, перейти к п. 2.2 алгоритма.

3. Если $|alpha|\neq0$, то вычислить функцию подобия $F = s/|alpha|$, где $|alpha|$ – длина собранного в переменной alpha сценария.

В результате выполнения алгоритма в переменной *alpha* будет собран сценарий семантического объекта и вычислена степень его присутствия в текстовом потоке.

Следует отметить, что п. 2.4 алгоритма задает так называемый бектрекинг– откат на шаг назад, когда не найден вывод части сценария, собранного в переменной *alpha*. В этом случае выбранная характеристика отбрасывается и выполняется выбор новой из непомеченных характеристик множества *P*.

Формально-грамматический подход обладает хорошими точностными характеристиками и не требует обучения. Для его реализации необходима только формальная модель семантического объекта, которая при необходимости может быть уточнена. Однако трудоемкость алгоритма сильно зависит от вида выбранной формальной грамматики, которая задает процедуру вывода $Z\Rightarrow+alpha$ [16, с. 202–339]. Вопрос подбора формальной грамматики для поиска быстрых решений требует отдельного обсуждения.

Формально-грамматический подход существенно упрощается и ускоряется, когда множество $L(P)$ является регулярным. Тогда грамматику можно представить регулярным выражением, значением которого будет множество $L(P)$, а вывод $Z\Rightarrow+alpha$ свести к последовательному разбору сценариев конечным распознавателем. Анализ естественно-языковых потоков показывает, что этот случай типичен на практике. Рассмотрим его.

Регулярное выражение представляет собой формулу алгебры, которая определяется аксиоматически через операции “|” – ИЛИ, “*” – катенации и “{}” – итерации следующим образом:

1. $\varnothing$ (пустое множество), λ (пустая цепочка) и имена лингвистических характеристик p_1, p_2,…, p_n – регулярные выражения по определению (аксиома, основание).

2. Если e_1 и e_2 – регулярные выражения, то $e_1 \mid e_2$ и $e_1 * e_2$ – также регулярные выражения. Обычно в регулярных выражениях знак операции * опускают и не используют.

3. Если e – регулярное выражение, то $\{e\}$ – также регулярное выражение.

4. Других регулярных выражений нет.

Значением регулярного выражения E является множество последовательностей символов (характеристик) из P, обозначаемое как $|E|$:

$$|E| = \left\{\alpha = p_{r1}p_{r2}\ldots p_{rk},\ p_{ri} \in P\ , \alpha \in L(P)\right\}. \quad (7)$$

Конечный распознаватель для последовательностей $|E|$ представляется системой переходов. Ее диаграмма состояний имеет одно начальное состояние S, одно заключительное состояние Z, которые соединены обобщенной дугой E. Дуге E придается следующий смысл – любая цепочка $\alpha_i \in L(P)$ переводит распознаватель из состояния S в состояние Z. При конструировании распознавателя систему переходов путем декомпозиции (разложения формулы E на подформулы) приводят к виду, когда дуги представляются характеристиками p_i и пустой цепочкой λ, такую систему переходов называют приведенной. Разбор сценария $\alpha_i \in L(P)$ приведенной системой переходов происходит последовательно по образующим его характеристикам, что эквивалентно поиску пути из начального состояния S в конечное состояние Z. В работе [17, с. 124–162] рассмотрены теоретические основы конструирования конечных автоматов подобного рода.

Модифицируем описанный выше алгоритм выявления семантического объекта с учетом регулярности языка сценариев $L(P)$. Пусть нормализованный текстовый поток представлен последовательностью токенов $T = t_1t_2\ldots t_l$, $P = \{p_1, p_2,\ldots, p_n\}$ – характеристическое множество семантического объекта, $L(P)$ – множество его сценариев и распознаватель сценариев представлен приведенной системой переходов, для которой множество Q – множество состояний распознавателя. Функции текстовой переменной *alpha* и переменной *s* остаются без изменений. Введем дополнительно стек состояний *St*, в котором будут накапливаться состояния распознавателя из Q, входящие в найденный путь, и определим для него операции ← (втолкнуть) и → (вытолкнуть). Для распознавателя также определим функцию переходов вида $q_t = M(q,p_j)$, которая по текущему состоянию q_j, и характеристике p_t системы переходов определяет новое состояние q_t. Текущее состояние распознавателя будем сохранять во вспомогательной переменной *q*.

Алгоритм:

1. Положить начальные значения
$alpha = \lambda$, $s = 0$, $St = St \leftarrow S$.

2. Выполнять последовательно для каждого токена t_i, $i = 1, 2, \ldots l$ текстового потока $T = t_1t_2\ldots t_l$ пока $S_k \neq Z$:

2.1. Снять метки использования с характеристик множества $P = \{p_1, p_2,\ldots, p_n\}$.

2.2. Выбрать из непомеченных характеристик множества P характеристику p_j с максимальным и ненулевым значением $w_j = Cprox(p_j,t_i)$ и пометить характеристику p_j как использованную.

2.3. IF переход $q = M(St, p_j)$ существует
THEN
begin

2.3.1. $St := St \leftarrow q$;

2.3.2. Выполнить катенацию переменной *alpha* с характеристикой p_j вида $alpha := alpha+p_j$;

2.3.3. увеличить переменную $s := s+w_i$;

2.3.4. $q := \text{Сост}\left(St, p_j\right)$
end
ELSE перейти к п. 2.2.

Вычислить функцию подобия $F = s/|alpha|$.

Отметим, что данный алгоритм работает намного быстрее, поскольку разбор сценария сводится к вычислению функции переходов. Основную трудоемкость алгоритма определяет только трудоемкость вычисления семантических близостей характеристик и токенов.

Предложенный формальный подход реализован в разработанном программном комплексе, включающем подсистему подготовки текстовых потоков, обеспечивающую их форматирование и токенизацию, подсистему создания и редактирования семантических объектов, подсистему поиска и выявления семантических следов объектов и подсистему сравнения токенов на семантическую близость.

Подсистема подготовки обеспечивает приведение к единому формату текстовых потоков от различных источников и их токенизацию. Подсистема создания и редактирования семантических объектов в полуавтоматическом режиме позволяет анализировать представленные внешние неформализованные исходные данные о семантическом объекте, в которых отражены характерные черты и особенности поведения семантических объектов, и создавать по ним их формальные описания (характеристические множества, регулярные выражения и системы переходов). Формальные описания заносятся в базу данных семантических объектов, из которой передаются в другие подсистемы в формате JSON файлов. Подсистемы семантического сравнения токенов и выявления семантических следов семантических объектов реализуют основной функционал комплекса и обеспечивают поиск и идентификацию семантических следов в текстовых потоках, включая определение степени доверия к результатам.

Для тестирования программного комплекса и проведения экспериментов была разработана следующая реалистично подобная схема. Некий субъект К, обладая информацией о нерешенных проблемах некоего Ж и преследуя собственную цель, вступил с ним в чат-диалог. В процессе диалога К завоевывал доверие Ж, проявляя сочувствие к его проблемам, вовлекал в игровой тренинг, предлагая помощь, и на завершающем этапе требовал неукоснительного выполнения всех указаний. Диалог адаптировался под конкретную ситуацию и мог происходить по различным сценариям.

В данной схеме К моделировался семантическим объектом, чат соответствовал текстовому потоку. Программный комплекс по семантическим следам идентифицировал объект и оценивал степень его присутствия в потоке. В ходе экспериментов изучалось влияние точности описания семантического объекта на качество распознавания: описание уточнялось за счет добавления новых характеристик и расширения языка сценариев. Результаты показали, что точность идентификации росла с повышением детализации описания, однако после достижения определенного порога дальнейшее уточнение не приводило к значимому улучшению. Для повышения достоверности результатов генерировались

разнородные диалоговые потоки, и исследования подтвердили, что точность идентификации не зависела ни от размера потока, ни от его вида. Дополнительно результаты идентификации и исходные данные оценивались экспертной группой, чьи выводы согласовывались с полученными данными. Результаты тестовых испытаний показали работоспособность программного комплекса и подтвердили основные теоретические положения исследования.

Результаты исследования и их обсуждение

Основной результат исследования состоит в разработке нового формального подхода к выявлению текстов определенной семантической направленности по описаниям их источников. Для этого вербальное понятие «определенная семантическая направленность» моделируется множеством сценариев языка формальной грамматики некоторого гипотетического семантического объекта, сценарии представляются последовательностями характеристик семантического словаря, а направленность текста определяет семантическая близость сценарию. Трудность реализации подхода кроется в бесконечной множественности сценариев, а также в отсутствии информации о семантической направленности исходного текста. В таких условиях простой перебор невозможен и предложено предполагаемый сценарий конструировать.

Процесс конструирования организуется последовательным определением семантического сходства токенов текста характеристикам и их сборкой в последовательность, которая путем построения вывода в формальной грамматике, называемым разбором, проверяется на принадлежность языку сценариев. При положительном исходе разбора сценарий построен.

Степень семантического сходства текста и сценария определяется значением специально сконструированной функции семантического подобия.

Для реализации подхода разработаны общий и частный алгоритмы выявления текстов определенной семантической направленности. В общем алгоритме разбор сводится к построению вывода в формальной грамматике, для регулярных грамматик разбор выполняется системой переходов. Ускорение процесса достигается путем совмещения сборки сценария с грамматическим разбором, а также использованием механизма бэк-трекинга. Отличительная особенность состоит в том, что точность алгоритмов определяется только фактической близостью текстов сценариев.

Заключение

Новые задачи, поставленные практикой естественно-языковой обработки, в основе которых лежит выявление текстов определенной семантической направленности по формальному описанию их источников, потребовали разработки нового теоретического инструментария – методов, моделей и эффективных алгоритмов и программных реализаций. В предлагаемой работе такой подход разработан на основе положений вычислительной теории семантической интерпретации и теории формальных грамматик. В нем использованы понятия вычислительной теории семантической интерпретации – семантическое сравнение, критерий семантической близости. Отличительная черта предлагаемого подхода состоит в том, что его точность определяется вычисляемой степенью фактической семантической близости сценариев и текстов. Поход не требует процедур обучения, как это имеет место в нейросетевом подходе. Следует отметить, что особенности разработки программ и результаты экспериментов в работе упомянуты кратко, поскольку данный вопрос требует отдельного обсуждения.

Работа направлена на развитие фундаментальных основ математического моделирования фундаментальных и прикладных проблем естественно-языковой обработки для создания эффективных вычислительных алгоритмов виде комплексов проблемно-ориентированных программ на основе современного компьютерного инструментария.

Список литературы

1. Прохоренок Н.А., Дронов В.А. Python 3. Самое необходимое. СПб.: БХВ-Петербург, 2021. 605 с. [Электронный ресурс]. URL: https://rusneb.ru/catalog/000200_000018_RU_NLR_BIBL_A_012485392/ (дата обращения: 16.02.2025).

2. Бенгфорт Б., Билбро Р., Охеда Т. Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка. СПб.: Питер, 2019. 368 с. [Электронный ресурс]. URL: https://rusneb.ru/catalog/000200_000018_RU_NLR_BIBL_A_011955723/ (дата обращения: 16.02.2025).

3. Николаев И.С., Митренина О.В., Ландо Т.М. Компьютерная и прикладная лингвистика. М.: Ленанд, 2016. 320 с. [Электронный ресурс]. URL: https://pureportal.spbu.ru/ru/publications/---(fa123ecb-1a1c-4ee2-b364-8351fabfff6c)/export.html (дата обращения: 16.02.2025).

4. Devlin Jacob, Ming-Wei Chang, Kenton Lee, Toutanova Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings – 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL. 2019. Vol. 1 (Long and Short Papers), P. 4171–4186. DOI: 10.18653/v1/N19-1423.

5. Azhar Kassem Flayeh, Yaser Issam Hamodi, Nashwan Dheyaa Zaki. Text Analysis Based on Natural Language Processing (NLP) // Proceedings – 2nd International Conference on Advances in Engineering Science and Technology, AEST. 2022. P. 774–778. DOI: 10.1109/aest55805.2022.10413039.

6. Вишняков Ю.М., Вишняков Р.Ю. Формализация распознавания и идентификации семантических объектов в естественно-языковых текстовых потоках // Известия ЮФУ. Технические науки. 2024. № 4. С. 110–122. URL: https://izv-tn.tti.sfedu.ru/index.php/izv_tn/article/view/985 (дата обращения: 16.02.2025).

7. Kai Hu, Huayi Wu, Kunlun Qi, Jingmin Yu, Siluo Yang, Tianxing Yu, Jie Zheng, and Bo Liu. A domain keyword analysis approach extending Term Frequency-Keyword Active Index with Google Word2Vec model // Scientometrics 114. March 2017. Vol. 114 (3). P. 1031–1068. DOI: 10.1007/s11192-017-2574-9.

8. Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, Siena Dumas Ang. Parsing algebraic word problems into equations // Transactions of the Association for Computational Linguistics (TACL). 2015. Vol. 3. P. 585–597. DOI: 10.1162/tacl_a_00160.

9. Chandrapati L.M., Rao C.K. Descriptive Answers Evaluation Using Natural Language Processing // IEEE Access. June 2024. Vol. 12. P. 87333–87347. DOI: 10.1109/ACCESS.2024.3417706.

10. Aman Bhadouria, Pranav Gupta, Parish Bindal, Kapil Madan, Sonal Sonal. Automated Examination System using Machine Learning and Natural Language Processing // Proceedings – 2024 Sixteenth International Conference on Contemporary Computing, IC3 2024. 2024. P. 752–761. DOI: 10.1145/3675888.3676144.

11. Tianshuo Peng, Zuchao Li, Lefei Zhang, Hai Zhao, Ping Wang, Bo Du. Multi-modal Auto-regressive Modeling via Visual Tokens // Proceedings – 32nd ACM International Conference on Multimedia, MM'24. 2024. P. 10735–10744. DOI: 10.1145/3664647.3681685.

12. Modi A., Dhanjal Y.S., Larhgotra A. Semantic Similarity for Text Comparison between Textual Documents or Sentences // Proceedings – 2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems, ICSES 2023. 2023. P. 1–5. DOI: 10.1109/ICSES60034.2023.10465440.

13. Sonali Mhatre, Shilpa Satre, Mansi Hajare, Aditi Hire, Aniket Itankar, Shruti Patil. Text Comparison Based on Semantic Similarity // Proceedings – 2023 3rd International Conference on Intelligent Technologies, CONIT 2023. P. 1–5. DOI: 10.1109/CONIT59222.2023.10205616.

14. Aadeesh Bali, Aniket Bhagwat, Aditya Bhise, Sarang Joshi. Semantic Similarity Detection and Analysis for Text Documents // Proceedings – 2024 Second International Conference on Emerging Trends in Information Technology and Engineering, ICETITE 2024. P. 1–9. DOI: 10.1109/icETITE58242.2024.10493834.

15. Вишняков Ю.М., Вишняков Р.Ю. Вычислительная семантическая интерпретация текстов научно-технического стиля // Современные наукоемкие технологии. 2016. № 12–2. С. 236–242.; URL: https://top-technologies.ru/ru/article/view?id = 36428&ysclid = m77urlmrjv786427462 (дата обращения: 16.02.2025).

16. Льюис Ф., Розенкранц Д., Стирнз Р. Теоретические основы проектирования компиляторов / Пер. с англ. В.А. Исаева и др. Под ред. В.Н. Агафонова. М.: Мир, 1979. 645 с. [Электронный ресурс]. URL: https://rusneb.ru/catalog/000199_000009_007626193/?ysclid = m77v68mwqf194123846 (дата обращения: 16.02.2025).

17. Ахо А., Ульман Дж. Теория синтаксического анализа, перевода и компиляции / Пер. с англ. В.Н. Агафонова. Под ред. В.М. Курочкина. М.: Мир, 1978. Т. 1. 612 с. [Электронный ресурс]. URL: https://rusneb.ru/catalog/000199_000009_007597729/ (дата обращения: 16.02.2025).