ФОРМАЛИЗАЦИЯ ПРОБЛЕМЫ ИЗВЛЕЧЕНИЯ ЗНАНИЙ ИЗ ЕСТЕСТВЕННО ЯЗЫКОВЫХ ТЕКСТОВ

Александр Палагин, Сергей Крывый, Николай Петренко, Дмитрий Бибиков

***Abstract**: Рассматривается формализация процесса анализа естественно-языковых текстов с целью извлечения знаний. Предлагается автоматизированный итеративный подход к реализации такого анализа.*

***Keywords**: автоматизация обработки ЕЯТ, извлечение знаний, онтограф текстового документа.*

***ACM Classification Keywords**: I.2 ARTIFICIAL INTELLIGENCE – I.2.4 Knowledge Representation Formalisms and Methods.*

Введение

Проблема извлечения знаний из естественно-языковых текстов (ЕЯТ) является одной из главных проблем в исследованиях по искусственному интеллекту. Этой проблеме в последнее время уделяется большое внимание в основном из-за того, что потоки информации неуклонно возрастают и человек уже не в состоянии самостоятельно обрабатывать эту информацию. Это относится в первую очередь к информации, которая находится в текстах книг, разного рода электронных коллекциях, статей, газетах, Интернете и т. п. В связи с таким положением дел возникает необходимость в разработке средств автоматизации для анализа прежде всего ЕЯТ на предмет извлечения из них релевантной запросу пользователя информации или убедиться с их помощью, что такой информации в этих текстах нет. Создание таких средств наталкивается на сложность проблемы анализа, которая в свою очередь связана с семантической многозначностью значений предложений естественного языка. Кроме этой многозначности существует еще ряд вопросов, связанных с анализом эмоциональной окраски фраз, анализом фраз иронического и иносказательного характера, анализом при неполной информации (подразумевается нечто по умолчанию или вообще неизвестно) и т. п. Поскольку такого типа проблемы плохо формализуются, то отсюда и вытекают причины огромной сложности проведения такого анализа.

Краткая история проблемы

Проблема автоматизации процесса анализа ЕЯТ с целью извлечения знаний занимала многих исследователей. На первый план здесь выходит проблема формализации семантики естественного языка (ЕЯ) и такая попытка была предпринята еще в начале 30-х годов прошлого столетия в работах А. Тарского и его учеников. Однако, о такой необходимости говорили еще Аристотель, Лейбниц, Эйлер и др. В частности, Аристотель выделил четыре типа высказываний, которые были названы силлогизмами, а сам подход Аристотеля был назван силлогистикой [1, 7, 16]. Позже Эйлер изложил свое понимание силлогистики Аристотеля с помощью геометрической интерпретации его силлогизмов [16] (эту интерпретацию стали называть кругами Ейлера). Далее идеи Эйлера были развиты в работах

французского математика и астронома Ж.Д. Жергона, который ввел типы отношений и интерпретацию силлогистики Аристотеля в терминах этих отношений [19]. Он показал, что каждый тип силлогизма Аристотеля можно выразить в виде некоторых возможных вариантов таких отношений. Главная трудность в использовании жергоновых отношений состояла в том, что практически все их типы в сложном предложении требовали анализа большого числа вариантов.

Более существенные шаги на пути формализации ЕЯ были сделаны А. Тарским [2, 3], в результате которых появилось понятие выполнимости формул – понятия более общего, чем понятие истинности. Это понятие Тарский применил к открытым и замкнутым формулам (под замкнутой формулой в предложениях ЕЯ понимают фразу), что позволило сформулировать понятие истинности предложения ЕЯ и наложить ограничения на каждую открытую атомарную формулу, которая состоит из атомарного предиката. Поскольку таких формул имеется только конечное множество, то такой подход становится конструктивным.

Следующая попытка улучшения формализации А. Тарского, была предпринята Д. Дэвидсоном [4]. Он предложил добавить к понятию выполнимости и истины рекурсивное определение истины. В таком случае теория Т, которая включает рекурсивное определение истины, объясняет каким образом значения фраз зависит от значения (интерпретации) слов в этих фразах.

Монтегю [5] тоже верил в то, что методы формальной семантики можно применить к исследованию семантики ЕЯ. Но он, в отличии от Дэвидсона, отказался от применения логики предикатов первого порядка и предпочел категориальные грамматики. Эта грамматика включает в себя те категории, которые специалисты в области грамматик традиционно используют при определении ЕЯ, например, такие категории, как прилагательное или причастие. Это позволило Монтегю заменить понятие абсолютной истины понятием относительной истины в модели, потому что в разных моделях одно и тоже предложение может иметь разные значения истинности. Такое расширение дало возможность определить понятие логической истинности и логического следствия для более широкого фрагмента ЕЯ. Таким образом, Монтегю выделил два элемента: интенцию (смысл) и экстенцию (денотат) и применил их к существительным, прилагательным и фразам. Существуют и другие подходы к анализу ЕЯ, которые базируются на понятиях семантической сети, фрейма и т. п. [18].

Недостатки и достоинства применения языка формальной логики к анализу предложений ЕЯ описаны в книгах [7] и [17], где исследуются рамки применения языков формальной логики к анализу предложений ЕЯ и указываются некоторые причины того, почему эти языки не всегда применимы к такого рода анализу. А результатом анализа предложений ЕЯ являются рассуждения (высказывания). Эти причины приводят к необходимости введения понятия достоверного и правдоподобного рассуждения (понятие правдоподобного рассуждения ввел Д. Пойа [6] и обосновал необходимость его использования в математике). Здесь под рассуждением понимается построение последовательности фактов, которые неизбежно приводят к принятию некоторого утверждения, являющегося целью рассуждения. Отметим, что понятие рассуждения существенно отличается от понятия логического вывода хотя бы тем, что оно может опираться на нелогические или металогические понятия как достоверного вывода, так и правдоподобного вывода.

Для последних десятилетий 20-го столетия в области исследования семантики ЕЯ характерными были попытки построения формальной теории семантики, которая была бы общей для естественных и

искусственных языков. Синтаксисом интересовались только в связи с семантикой, основной целью которой являлось объяснение понятий истины и логического следствия, а основной целью синтаксиса была характеристика синтаксических категорий, из которых строятся предложения.

Что касается данной работы, то она является продолжением исследований, начатых в работах [9-14] и связанных с проблемой анализа предложений ЕЯ с целью извлечения знаний.

Характеристика проблемы

Пусть заданы X – алфавит некоторого естественного языка, а $F(X)$ означает множество слов в алфавите X. Рассмотрим $L \subseteq F(X)$ – естественный язык в данном алфавите, предложения которого построены в соответствии с правилами грамматики P, где $P = \{p_i : i = 1,\ldots,m\}$ – правила грамматики языка L. Правила грамматики определяют совокупность отношений

$$R_E = \left\{ R_{p_i} : p_i \in P \right\},$$

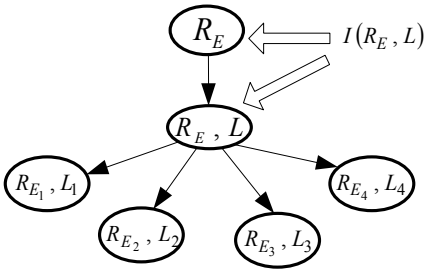
каждое из которых соответствует конкретному правилу грамматики. Пусть слова языка L разбиты в словаре этого языка на лексико-грамматические разряды с помощью лексико-грамматического отношения R [20–22]. Это означает, что в один класс попадают существительные, в другой класс – глаголы, в третий – прилагательные и т. д. Очевидно, что это отношение является отношением эквивалентности и в соответствии с этим отношением слова из L разбиваются на классы, элементы которых соответствуют лексико-грамматическим разрядам языка L, т. е. $L = L_1 \cup \ldots \cup L_j \cup \ldots \cup L_k$ – конечное множество лексико-грамматических разрядов в языке L.

Пусть L_i – некоторый класс этого разбиения. Слова, которые входят в L_i структурируются в соответствии с лингвистическими и семантическими отношениями языка L, являющимися отношениями частичного порядка или квазипорядка (например, гипоним-гипероним, мероним-голоним или род-вид, целое-часть, класс-элемент, вышестоящий-нижестоящий, класс-подкласс). Отношения частичного порядка задаются в виде ориентированного или неориентированного графа $G = (V, E)$, где $V = L_i$, а E – множество пар слов (p,q), $p,q \in L_i$ таких, что p доминирует над q по одному из указанных отношений. Такой граф называют онтографом (в математике такие графы называются диаграммами Хассе).

Заметим, что хотя в определении вершин онтографа использовались разряды $V = L_i$, на самом деле вершинами являются классы эквивалентных между собой слов из L_i относительно глобального отношения синонимии RS. Таким образом под $V = L_i$ следует понимать фактор множество $V = L_i / R_S$. В таком понимании онтограф $G = (V, E)$ представляет собой гиперграф, вершины которого соответствуют классам синонимичных слов языка. Далее под онтографом будем понимать именно такого типа гиперграф.

При конкретизация языка L и проблемы, которая исследуется, выполняется конкретизация отношений и построение соответствующего онтографа для данного естественного (украинского, русского, английского)

языка. Чаще всего в работах на эту тему рассматриваются следующие четыре разряда языка L: L1 – существительные, L2 – глаголы, L3 – прилагательные и L4 – наречия. В соответствии с такой семантической интерпретацией классов L_i как лексико-грамматических разрядов существительного, глагола, прилагательного и наречия ($\{L_1, L_2, L_3, L_4\}$) онтограф языково-онтологической картины мира (ЯОКМ) строится в соответствии со следующей схемой:



В данной работе нас будут интересовать отношения из $R_E = \{R_{p_i} : p_i \in P\}$ и их интерпретация $I(R_E, L)$. Отношения из R_E определяют синтаксические правила построения предложений языка *L,* но при анализе предложений такого языка на первый план выступают семантические отношения $I(R_E, L)$, как интерпретация отношений из R_E, поскольку синтаксически правильные предложения могут быть абсолютно бессмысленными с точки зрения здравого смысла.

Возникает вопрос: как определить семантические отношения $I(R_E, L)$ на синтаксически правильных предложениях языка *L*, т.е. предложениях, принадлежащих к отношениям из R_E?

В связи с тем, что язык является естественным, то однозначного способа определения семантики, т.е. интерпретации предложений этого языка, дать невозможно. По видимому в этом и состоит главная трудность в решении проблемы анализа предложений естественного языка с целью извлечения знаний. Это обстоятельство дает определенную свободу в определении такой семантики, которая налагает определенные обязательства. Мы примем следующее определение семантики.

Определение 1. Синтаксически правильное предложение s языка L будем называть семантически правильным, если оно является фактически истинным, или достоверным, или правдоподобным с точки зрения здравого смысла группы индивидуумов или отдельно взятого индивидуума.

Понятно, что приведенное определение, не является строгим, но оно, в определенной степени, соответствует естественному положению дел. Приведем краткие комментарии понятий, фигурирующих в этом определении.

Под ***фактически истинным*** предложением понимается умозаключение, факты которого адекватно интерпретируются (т.е. имеют единственно возможное логическое значение) в языке формальной логики (логики высказываний, предикатов, модальной и т.п.) или получено из таких фактов по одному из формальных правил вывода в этой логике. Если удается извлечь фактически истинные знания из предложений ЕЯ, то это позволяет полностью решить логические проблемы анализа.

Пример 1. Примерами фактически истинных предложений являются следующие высказывания: «*Земля круглая и вращается вокруг Солнца», «Г.Ф. Вороной – украинский математик», «число 23 – простое».*

А для иллюстрации анализа такого типа предложений, рассмотрим следующий текст: *«Жители сельской местности будут голосовать за кандидатуру президента (ГП) только в том случае, если он подпишет закон о праве сельских жителей на землю (ЗЗ). Рабочие заводов и олигархи (РО) не будут голосовать за кандидатуру президента, если он не наложит вето на этот закон (ВЗ). Очевидно, что президент либо подпишет закон или наложит на него вето. Следовательно, он либо потеряет голоса жителей сельских регионов, либо голоса рабочих и олигархов».*

Анализ данного фрагмента текста допускает следующую интерпретацию в языке логики высказываний:

Дано: $ГП \rightarrow ЗЗ$, $РО \rightarrow ВЗ$, $\neg ЗЗ \vee \neg ВЗ$. Верно ли, что из этих фактов следует $\neg ГП \vee \neg РО$.

Методом резолюций справедливость этого следствия доказывается в два шага. $\triangle$

Под ***достоверным предложением*** понимается такой факт, который подтверждается имеющимся опытом или знаниями либо группы людей либо отдельного взятого человека или следует из кажущихся им истинных предположений по какому-нибудь правилу умозаключения. При исследовании на непротиворечивость умозаключений, извлеченных из достоверных предложений, в широком смысле наиболее характерен дедуктивный вывод. Отсюда следует, что фактические (логические) следствия в языке формальной логики являются очень частным случаем достоверных умозаключений. Следует заметить, что дедукция «в высшей степени идеализированная и ограниченная форма рассуждений» [8], которая применима только в очень узких рамках к моделированию и анализу предложений ЕЯ, поскольку не полностью отражает такие понятия как здравый смысл, неопределенность, достоверность информации и т. п.

Далее, понимание разными людьми смысла одного и того предложения может быть разным, а отсюда вытекает, что точку зрения группы людей или отдельно взятого человека необходимо учитывать (кстати, разное понимание одного и того же предложения (высказывания, рассуждения) является источником возникновения *дискуссии*). В действительности мы очень редко пользуемся абсолютно достоверными фактами, поскольку многообразие мира, описываемое этими фактами, нельзя ограничить и втиснуть в какие-нибудь формальные рамки. Поэтому в процессе рассуждений в большинстве случаев мы оперируем, *опираясь на правдоподобные факты, а не на достоверные факты*. Это связано с тем, что часто при принятии решения о чем-нибудь, мы прибегаем к наблюдению и опыту. А это нам дает только правдоподобные факты, которые потом должны проверяться и доказываться, и только после этого приниматься или опровергаться.

Под ***правдоподобным предложением*** будем понимать такой факт, истинность которого опирается на кажущиеся правильными (истинными) умозаключения, с точки зрения группы людей или отдельно взятого человека. Для правдоподобных рассуждений характерным является индуктивный вывод и вывод по аналогии [6].

Отметим также, что многие исследователи этой проблемы также считают, что моделирование знаний, извлеченных из ЕЯТ, не может ограничиваться формализацией только лишь непогрешимого интеллекта. Естественным основанием для такого мнения является то, что важной чертой естественного интеллекта есть способность вырабатывать здравые рассуждения, которые могут оказаться и недостоверными. В случае неполной, неточной и изменчивой информации наши рассуждения часто становятся только

предположительными, а в следствии этого лишь правдоподобными и поэтому могут подлежать пересмотру и уточнению (модификации).

Другие исследователи считают, что проблема анализа ЕЯТ решена, если извлеченные знания представлены в базе знаний и все проблемы по их анализу решаются средствами баз знаний. Это мнение, с нашей точки зрения, не совсем соответствует реальному состоянию дел. Главной проблемой при обработке знаний в базах знаний является та же модифицируемость знаний. Модификация знаний необходима по многим причинам. В частности, в самой базе знаний объекты могут представляться интенсионально (аксиоматически) или экстенсионально (перечнем элементов), что вносит свои коррективы в процесс их обработки. А в общем случае различают два основных типа модифицируемых знаний внешнего характера: *предположительные и предполагаемо полные.*

Предположительные знания являются всего лишь правдоподобными. Это связано с тем, что они неточные, поскольку базируются на неполной, неточной и изменчивой информации, а также по причине их естественной неточности и модифицируемости. Примерами такого типа знаний являются рассуждения по умолчанию, рассуждения с прототипами и знания статистического характера. Неполнота знаний является естественной, поскольку в повседневной жизни мы часто общаемся путем молчаливо подразумеваемых фактов. Такого типа факты в силлогистике именуются *антимемами*. Антимемы неизбежны потому, что они существенно ускоряют процесс обмена мыслями между людьми и без них этот процесс сделался бы невыносимо скучным.

Пример 2. Рассмотрим такой текст: «Уважаемая госпожа Хадсон, если я за ужином выпью крепкий напиток вроде виски (ВВ), то я не смогу заснуть (З) всю ночь. Поэтому, с Вашего разрешения, я за ужином не буду пить виски.».

Это пример антимемы, которая выглядит таким образом: из $ВВ \to \neg З$ вытекает $\neg ВВ$. Очевидно, что пополненное пропущенным фактом рассуждение выглядит так: из $ВВ \to \neg З$, $З$ вытекает $\neg ВВ$. $\triangle$

Предполагаемо полные знания – это знания, которые основаны на фактах, которые предполагаются информационно полными, но которые таковыми не являются или перестают быть таковыми. В действительности часто встречается ситуация, когда выдвигаются модифицируемые (а иногда и неявные) соглашения, для наращивания наших знаний в условиях неполной или неизвестной информации. Основываясь на таких знаниях наши выводы могут быть логически корректными по отношению к этим добавленным знаниям. Однако, эти рассуждения оказываются модифицируемыми, так как они основываются на изменчивом состоянии знаний. Следует отметить, что некоторые корректные формы вывода, которые формализует классическая математическая логика, также могут оказаться модифицируемыми. Это объясняется тем, что они применяются к базе знаний, которая зачастую пополняется всего лишь правдоподобными знаниями. Например, знания, занесенные одним исследователем в базу знаний, могут быть неправильно или неточно поняты другими и поэтому будут подвергаться модификации другими исследователями.

Формальная постановка задачи

В связи с проблемой, которая нас интересует, необходимо определить понятия «знание» и «процесс извлечения знаний» из предложений ЕЯ[1]. Строгого определения понятия «знание» не существует, однако это понятие вызывало большой интерес ученых, начиная с древних греков. Его изучали Платон и Аристотель, которые ввели еще целый ряд понятий, характеризующих знание: «рассудок», «мнение», «математическое мышление» и др. В 20-м столетии в связи с развитием такой области как программирование появились понятия «процедурное» и «декларативное» знание. Процедурное знание содержит в себе информацию о том, как нужно действовать, чтобы получить нужный результат, а декларативное знание содержит в себе информацию о том, над чем надо выполнять эти действия. В частности, в области искусственного интеллекта закрепилось и употребляется следующее определение понятия «знание». ***Знание – это обоснованное истинное убеждение (вера).*** Это определение для решения нашей задачи мало что дает, поскольку не очерчивает материальный объект.

В целях более точной формулировки понятий «знание» и «извлечение знаний», которыми будем пользоваться в этой работе, рассмотрим следующие определения, пользуясь нотацией констрейнтного программирования [15].

Пусть дано некоторое множество D, на котором определена конечная совокупность $R = \{R_1,\ldots,R_k\}$ отношений $R_i \subseteq D^n$, $i = 1,2,\ldots,k$, конечной арности. Языком ограничений L на D называется непустое множество $L \subseteq R$. Проблема выполнимости ограничений из L формулируется следующим образом.

Определение 2. Для произвольного множества D и языка ограничений L на D проблемой выполнимости ограничений $CSP(L)$ является решение такой комбинаторной задачи:

дана тройка $P = (V,D,C)$, где

– $V = \{v_1,\ldots,v_m\}$ - конечное множество переменных;

– $C = \{c_1,\ldots,c_q\}$ - конечное множество ограничений, где ограничение c_i из C - пара (s_i,R_i), где $s_i = (v_{i1},\ldots,v_{ij})$ - кортеж, состоящий из переменных, $R_i \in L - n_j$ -арное отношение на D;

найти функцию $\varphi : V \to D$ такую, что $\forall (s_i,R_i) \in C$ кортеж $(\varphi(v_{i1}),\ldots,\varphi(v_{ij})) \in R_i$ либо убедится в том, что ее не существует, $i = 1,2,\ldots,k$. Множество D в этом случае называется областью проблемы, а функция φ называется интерпретацией $CSP(L)$.

Применительно к анализу предложений ЕЯ с целью извлечения знаний множество D интерпретируется как множество объектов (сущностей), извлеченных из предложений входного текста *T*, удовлетворяющих отношениям из $R_E = \{ R_{p_i} : p_i \in P \}$, которое факторизовано по некоторому отношению эквивалентности R_S^s (это отношение представлено в онтографе вершинами синонимичных объектов, которые факторизуются по предметно-ориентированному отношению синонимии). Переменные из

[1] Один из вариантов таких определений мы привели в [14].

множества $V = \{v_1, v_2, \ldots, v_m\}$ принимают свои значения в этом факторизованном множестве объектов D, фигурирующих в тексте T (это могут быть более широкие лексико-грамматические разряды, такие, как конкретные личности, даты, конкретные предметы и т. п.). А в качестве $\varphi : V \to D$ выступает интерпретация $I(R_E, L)$, в результате которой появляются отношения (предикаты) $\{\phi_1, \phi_2, \ldots, \phi_m\}$.

Отношения $\{\phi_1, \phi_2, \ldots, \phi_m\}$ из $I(R_E, L)$, извлеченные из текста Т ЕЯ, будем называть **знаниями.**

Это определение, по нашему мнению, уточняет данное выше определение знания в том смысле, что оно материализует объект поиска и механизм этого поиска.

Пример 3. Пусть имеем такой текст: *«Гуляя набережной Черного моря писатель заметил, что облака (О) подобны пляшущим сатирам (С). Но присмотревшись, он понял, что эти сатиры плывут, а не пляшут».*

В этом тексте множество *D* ={набережная Черного моря (Н(м,ч)), писатель (П), облака (О), сатиры (С)} *R*={гулять, замечать, понимать, плыть, плясать, присматриваться, подобные}. В данном случае Н(м,ч) означает объект **набережная**, а **(м,ч)** – атрибуты (ограничения) этого объекта.

Из первого предложения получаем такие отношения:

гулять(П, Н(м,ч)), замечать(П,О), подобные(О,С), плясать(С).

Из второго предложения получаем такие отношения:

присматриваться(П,С), понимать(П), плыть(С), $\neg$плясать(С).

Отношение *подобные(О,С)* факторизует множество *D* на классы по объектно-ориентированному отношению синонимии. Заметим, что глобальное отношение синонимии не внесет объекты «сатиры» и «облака» в один класс синонимичных объектов, в то время как предметно-ориентированное отношение внесет объекты «облака» и «сатиры» в один класс синонимии (по отношению подобия). Таким образом, факторизованное множество *D* принимает вид: *D* ={Н(м,ч), (П), {О,С}}. Эта факторизация позволяет все высказывания относительно С считать аналогичными высказываниями относительно О и уточнить полученные отношения следующим образом:

гулять(П, Н(м,ч)) $\wedge$ *замечать(П,О), плясать(С)* $\leftrightarrow$ *плясать(О)*

для первого предложения и

(присматриваться(П,С) $\wedge$ понимать(П)) $\rightarrow$ (плыть(С) $\wedge$ $\neg$плясать(С))

для второго предложения.

Соединяя эти факты вместе и выполняя эквивалентные преобразования, получаем

гулять(П, Н(м,ч)) $\wedge$ *замечать(П,О),* *(присматриваться(П,С)* $\wedge$ *понимать(П))* $\rightarrow$ *плыть(С),*

(присматриваться(П,С) $\wedge$ *понимать(П))* $\rightarrow$ $\neg$*плясать(С),* *плыть(С)* $\leftrightarrow$ *плыть(О).*

В процессе логического анализа этих фактов (вследствие эквивалентных преобразований) возникает необходимость модификации извлеченных знаний.

Заметим, что этот текст описывает некоторое поэтическое (метафорическое) восприятие наблюдаемых фактов, которые в действительности не имеют места. $\triangle$

Автоматизированный итеративный метод анализа предложений ЕЯ

Исходя из выше сказанного, можно предложить такой итеративный способ автоматизированной обработки ЕЯТ [13].

Шаг 1. Морфологический анализ заданного текста T с целью построения словаря для текста T и разбиения на классы $\{L_1, L_2, L_3, L_4\}$ (или более мелкого разбиения, включающего и другие части речи). Кроме того, на этом шаге вычисляется парадигма всех словоформ изменяемых частей речи и исходная лексема, выделение отглагольных существительных и др.

Шаг 2. Построение множества объектов D, исходя из результатов синтаксического анализа текста T и результатов шага 1. Кроме того, на этом шаге вычисляются многословные термины, анафорические связи, антимемы и т. п.

Шаг 3. Построение онтографа, исходя из множества объектов D (построение отношения R_S) на классах $\{L_1, L_2, L_3, L_4\}$. Онтограф текста строится на основе онтографов предложений применением правил конъюнкции и упрощения, алгоритмы применения которых описаны в [14.]

Шаг 4. Построение интерпретации $I(R_E, L)$ на множестве объектов D, исходя из онтографа и предметно-ориентированного отношения синонимии на D.

Шаг 5. Внесение полученных на шаге 4 отношений $\{\phi_1, \phi_2, \ldots, \phi_m\}$ в базу знаний.

Шаг 6. Выполнить анализ множества отношений $\{\phi_1, \phi_2, \ldots, \phi_m\}$ средствами базы знаний.

Шаг 7. Если результаты анализа удовлетворяют пользователя, то закончить процесс иначе выполнить уточнение множества D и интерпретации $I(R_E, L)$ и перейти на шаг 3.

В приведенной последовательности шагов многие из них требуют комментариев. Заметим, что первые три шага детально изучались многими исследователями и для их реализации имеются соответствующие средства, работающие в автоматическом или автоматизированном режиме [10, 17, 20–22].

Наиболее проблемными являются шаги 4 и 7, что является следствием неформального определения семантически правильного предложения. На шаге 4 предполагается такая обобщенная схема взаимосвязей структурных компонент текста *T,* которая следует из семантической интерпретации соответствующих частей речи:

- объекты – это существительные;
- отношения (предикаты) – это глаголы;
- атрибуты объектов – это прилагательные (ограничения на объекты);
- атрибуты отношений (предикатов) – это наречия (ограничения на предикаты).

Такая интерпретация согласовывается с определением 2, со структурой предложений текста T и с другими известными концепциями (в частности, с концепцией системы WordNet).

Этот шаг, по-видимому, необходимо выполнять в автоматизированном (например, в диалоговом) режиме, консультируясь с пользователем, являющимся (или с пользователями являющимися) авторами текста T или экспертами в той предметной области, к которой относится данный текст T. На этом шаге сначала определяются имена отношений и их арности, которые связываются прежде всего с глаголами (как в вышеприведенном примере 3 имена отношений ПЛЯШУТ и ПЛЫВУТ и их арность 1). Затем, полученные

таким образом отношения, уточняются в процессе взаимодействия с пользователем. Если такое уточнение выполнено, то осуществляется переход на шаг 5.

Шаги 5 и 6 детально комментировать нет необходимости, поскольку представляется понятным, что на этих шагах должно выполнятся. Извлеченные из текста знания заносятся в базу знаний таким образом, чтобы можно было эффективным способом проводить их анализ. Выбор способа представления знаний в такой базе зависит от того, какие алгоритмы вывода будут использованы. Что касается такой обработки извлеченных из текста знаний в базе знаний, то о методах их представления, обработки и анализа свойств мы отсылаем читателя к монографии [8], где описаны основные методы и средства различных типов вывода.

Процесс выполнения шага 7 заключается в том, что если результаты анализа в базе знаний удовлетворяют пользователя, или подтверждают факты, полученные опытным путем, или соответствуют ожидаемым результатам, то дальнейший анализ можно не проводить и закончить работу алгоритма. В противном случае, если результаты анализа приводят к противоречиям, или являются абсурдными с точки зрения здравого смысла, или носят недостоверный характер, или не правдоподобны, то необходимо сделать повторный анализ семантических отношений, присутствующих в тексте, сделать необходимые уточнения или другие предположения и выполнить соответствующую модификацию, после чего повторить шаги 3 – 7. В процессе повторного анализа необходимо убедиться в правильности построенных семантических отношений, правильности дополнительных предположений, правильности трактовки некоторых понятий, объектов и отношений между этими объектами с точки зрения здравого смысла, если не удается эти отношения проинтерпретировать в языке математической логики.

Пример работы алгоритма

Рассмотрим пример работы первых трех шагов вышеприведенного алгоритма, поскольку продемонстрировать работу всех шагов нет возможности. С этой целью проанализируем такой текст (анализируемый текст написан на украинском языке, в связи с тем, что в данной версии в системе функционирует только толковый словарь украинского языка):

«В сучасному розумінні, комп'ютер – це універсальний електронний пристрій, призначений для автоматизації накопичення, збереження, опрацювання, передачі та відтворення даних.

Структурно комп'ютер складається з чотирьох основних пристроїв відповідно до тих завдань, які вони вирішують при опрацюванні даних.

Пристрої вводу призначені для вводу (накопичення) інформації та управління роботою комп'ютера користувачем. До цих пристроїв відносяться: клавіатура, маніпулятор миша, сканер, джойстик.

Пристрої виводу призначені для виводу інформації з метою візуального спостереження за роботою комп'ютера та створення твердих копій документів. До них належать монітор, принтер (друкуючий пристрій), плоттер.

Запам'ятовуючі пристрої призначені для збереження інформації, як тривалого, так і тимчасового, на час їх опрацювання комп'ютером. Пам'ять поділяють на внутрішню та зовнішню. Внутрішня призначена для збереження інформації під час роботи комп'ютера. Сюди відносять оперативну, постійну, кеш та CMOS-пам'ять. Вміст цієї пам'яті, як правило, зберігається лише при увімкненому живленні. Зовнішня пам'ять

призначена для тривалого збереження інформації незалежно від того, чи є живлення. Для зовнішньої пам'яті розрізнять пристрої пам'яті (накопичувачі) та носії даних (дискети, диски, магнітні стрічки, лазерні диски, тощо). Пристрій опрацювання інформації здійснює її переробку та загальне управління роботою всіх інших пристроїв. Цим пристроєм є процесор.

Головним пристроєм комп'ютера є центральний процесор. Він і виконує основні операції по опрацюванню даних та управління роботою інших пристроїв. По способу розташування пристроїв відносно процесора їх поділяють на внутрішні та зовнішні (периферійні). До внутрішніх відносіть деякі види пам'яті. Зовнішніми є пристрої вводу-виводу інформації, пристрої для її тривалого збереження. Узгодженість роботи окремих пристроїв здійснюють апаратні інтерфейси. Їх поділяють на послідовні та паралельні. Через послідовний інтерфейс дані передаються по одному біту. Вони прості за будовою, не вимагають синхронізації роботи передавача та приймача даних. Однак пропускна здатність їх менша, а коефіцієнт корисної дії нижчий. Їх використовують для підключення "повільних" пристроїв – наприклад, різноманітних датчиків».

Результат работы системы анализа этого текста на шаге 1 показан (частично) на рис. 1 (снимок экрана компьютера), где приведены элементы классов существительных (іменники), глаголов (дієслова), прилагательных (прикметники) и причастий (прислівники).

Статистика текста

Слова	Длинна	Частота	Часть речи
накопичення	11	2	іменник
опрацювання	11	3	іменник
відтворення	11	1	іменник
складається	11	1	дієслово
опрацюванні	11	1	іменник
відносяться	11	1	дієслово
маніпулятор	11	1	іменник
візуального	11	1	прикметник
тимчасового	11	1	прикметник
комп'ютером	11	1	Не распознано
увімкненому	11	1	дієприкметник
центральний	11	1	прикметник
опрацюванню	11	1	іменник
периферійні	11	1	прикметник
послідовний	11	1	прикметник
передаються	11	1	дієслово
збереження	10	5	іменник
структурно	10	1	прислівник
відповідно	10	1	прислівник
призначені	10	3	дієприкметник
інформації	10	7	іменник
управління	10	3	іменник

Рис. 1. Результат работы системы анализа текста на шаге 1

Далее, система также выдает номер предложения и номер вхождения данного слова в тексте. Эта информация используется на шаге 4 при построении и интерпретации предложений и извлечения

первичных знаний из текста и, в частности, при определении арностей извлекаемых предикатов (отношений).

Из приведенных рисунков видно, как очерчивается область интерпретации D и классы разбиения слов текста на классы $\{L_1, L_2, L_3, L_4\}$. В данном случае $L_1 =$ {…накопичення, опрацювання, маніпулятор, збереження, інформації, управління,…}, $L_2 =$ {…складається, відносяться, передаються,…}.

Система также выдает частоту появления слов в тексте и длину этих слов (рис. 2), что позволяет выполнить частотный анализ исходного текста с целью идентификации плагиата или чего-либо другого.

Кроме того, результат построения классов L_1 – L_6 для первых трех предложений анализируемого текста представлен в таблице ниже. А на рис. 3 – рис. 5 представлены онтографы предложений 1–4.

Для рассмотренного примера множество R_s состоит из элементов {R_1 (эквивалентность), R_2 (меронимгипероним), R_3 (назначение), R_4 (тема), R_5 (обрабатывать), R_6 (принимать решение)}.

Таблица Vt

Слова	№ вхождения	№ предложени
запам'ятовуючі	1	7
універсальний	1	1
автоматизації	1	1
спостереження	1	5
синхронізації	1	24
користувачем	1	3
CMOS-пам'ять	1	10
зберігається	1	11
накопичувачі	1	13
розташування	1	18
вводу-виводу	1	20
узгодженість	1	21
електронний	1	1
призначений	1	1
накопичення	1	1
накопичення	2	3
опрацювання	1	1

Рис. 2. Результат работы системы анализа текста на шаге 2

Заключение

Описанные в данной работе понятия и подходы к автоматизации обработки ЕЯТ составляют основу как теоретического, так практического анализа извлечения знаний из ЕЯТ. Данный подход используется в Институте кибернетики им. В.М. Глушкова НАН Украины и Киевском национальном университете им. Т. Шевченко в экспериментальной системе автоматизации анализа ЯЕТ с целью извлечения знаний в рамках онто-логического подхода к представлению и обработки информации. Используя эту основу и, прежде всего ее реализацию, предполагается наращивание ее мощности за счет построения новых метаотношений над построенными отношениями, являющимися отдельными частями знаний в исследуемом тексте, а также анализа текстов, написанных на русском языке.

Номер предложения	Класс L_1		Класс L_2		Класс L_3		Класс L_4 (наречие)	Класс L_5/ L_6 (причастие/ числительное)
	Исходное существ.-е	Лексема	Исходный глагол	Инфинитив	Исходное прилагательное	Лексема	Лексема	Лексема
1	розумінні комп'ютер пристрій автоматизації накопичення збереження опрацювання передачі відтворення даних	розуміння комп'ютер пристрій автоматизація накопичення збереження опрацювання передача відтворення дані	–	бути	сучасному універсальний електронний	сучасний універсальний електронний		призначений
2	комп'ютер пристроїв завдань опрацюванні даних	комп'ютер пристрій завдання опрацювання дані	склада-ється вирі-шують	склада-тися вирішу-вати	Основних	основний	струк-турно відпо-відно	/чотирьох (чотири)
3	пристрої вводу призначені вводу накопичення інформації управління роботою комп'ютера користувачем	пристрій ввід призначення ввід накопичення інформація управління робота комп'ютер користувач						
…	…	…	…	…	…	…	…	…

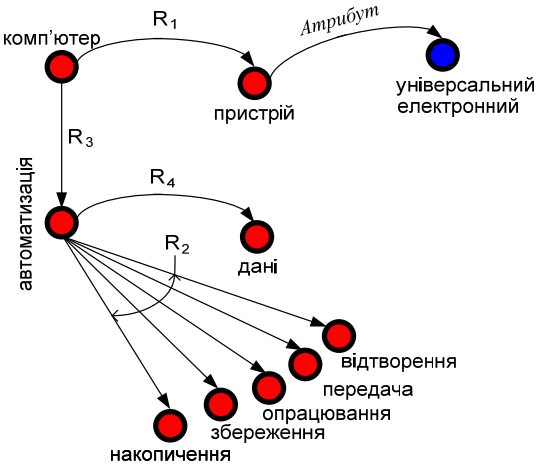


Рис. 3. Онтограф лексики предложения 1

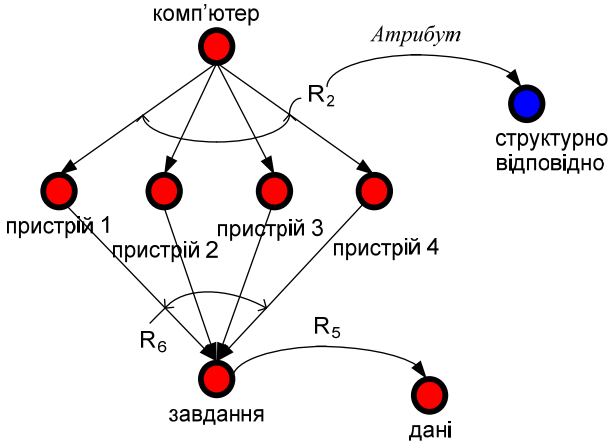


Рис. 4. Онтограф лексики предложения 2

Речення 3-4

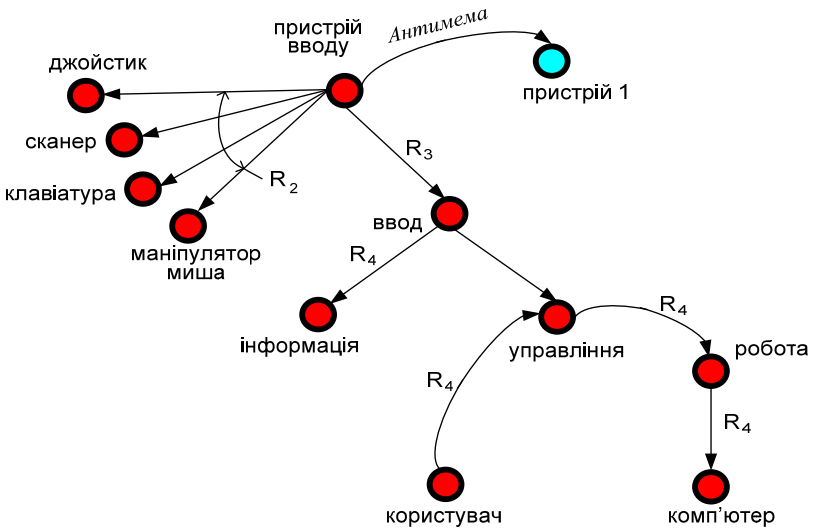


Рис. 5. Онтограф лексики предложений 3 и 4

Литература

1. Колмогоров А.Н., Драгалин А. Г. Введение в математическую логику. – Издательство Московского университета. – 1982. – 118 с.

2. Tarski A. The semantic conception of truth. Philosophy and phenomenological Research. – v.4. – 1944. – P. 241–375.

3. Tarski A. Logique, Semantique and Metamathematique (1923-1944). Colin. – Paris. – 1972.

4. Devidson D. Proceedings of Philosofical Logic. – Reidel. – Dordrecht. – 1969.

5. Montague R. Universal grammars. Theoria. Formal Phylisophy: Selected Papers of R. Montague. – Yale University Press.– 1974. – vol. 36. – P. 222–246.

6. Пойа Д. Математика и правдоподобные рассуждения. – М.: Наука.–1975.– 462 с.

7. Клини С. Математическая логика. – М.: Мир. – 1973. – 480 с.

8. Вагин В. А., Головина Е. Ю., Загорянская А. А., Фомина М. В. Достоверный и правдоподобный вывод. – М.: Физматлит. – 2004. – 703 с.

9. Палагин А. В., Крывый С. Л., Петренко Н. Г., Знание-ориентированные информационные системы с обработкой естественно-языковых объектов: основы методологии и архитектурно-структурная организация. – ж. УСиМ. – 2009. – №3. – С. 42–55.

10. Палагин А. В., Крывый С. Л., Бибиков Д. С. Обработка предложений естественного языка с использованием словарей и частоты появления слов. – Natural and Artificial Intelligence Intern. Book Series. – Inteligent Processing. – ITHEA. – Sofia. – N 9. – 2010. – P. 44–52.

11. Палагін О. В., Кривий С. Л., Петренко М. Г., Бібіков Д. С. Алгебро-логічний підхід до аналізу та обробки текстової інформації. – ж. «Проблемы программирования». – 2010. № 2–3. – С. 318–329.

12. Палагін О. В., Кривий С. Л., Бібіков Д. С., Величко В. Ю., К. Марков, К. Іванова, І. Мітов Формально-логічний підхід до побудови системи аналізу знань в різних предметних областях. – ж. «Проблемы программирования». – 2010. – № 2–3. – С. 382–389.

13. Крывый С. Л. Бибиков Д. С. Итеративный подход к анализу естественно-языковых текстов: логический аспект. – ж. «Проблемы программирования». – 2012. – № 2–3. – С. 318–329.

14. Палагин А. В. Онтологические методы и средства обработки предметных знаний / А. В. Палагин, С. Л. Крывый, Н. Г. Петренко. – [монография] – Луганск: изд-во ВНУ им. В. Даля, 2012. – 324 с.

15. Cohen D., Jeavons P. The Complexity of Constraint Languages. In “Handbook of Constraint Programming. – Edited by F. Rossi, P. van Beek and T. Walsh. – 2006. – P. 245 – 280.

16. Кулик Б. А. Логика естественных рассуждений. – С.-Петербург: Невский диалект. – 2001. – 127 с.

17. Рубашкин В.Ш. Представление и анализ смысла в информационных системах. – М.: Наука. – 1989. – 188 с.

18. Тейз А., Грибомон П., Луи Ж. и др. Логический подход к искусственному интеллекту. От классической логики к логическому программированию. – М.: Мир. – 1990. – 429 с.

19. Тейз А., Грибомон П., Юлен Г и др. Логический подход к искусственному интеллекту. От модальной логики к логике баз данных. – М.: Мир. – 1998. – 494 с.

20. Леонтьева Н.Н., Семенова С.Ю. Семантический словарь РУСЛАН как инструментарий компьютерного понимания. – М.: МГГИИ. – 2003. – С. 41–46.

21. Леонтьева Н.Н. К теории автоматического понимания естественных текстов. Часть 1. Моделирование системы "мягкого понимания" текста: информационно-лингвистическая модель. – М., МГУ, 2000. – 43 с.

22. Леонтьева Н.Н. К теории автоматического понимания естественных текстов. Часть 2. Семантические словари: состав, структура, методика создания. – М., МГУ, 2001. – 41 с.

Информация об авторах

Александр Палагин *– Ин-т кибернетики им. В.М. Глушкова НАН Украины, Киев-187 ГСП, 03680, просп. акад. Глушкова, 40; e-mail: palagin_a@ukr.net*

Область исследований: Общая теория знание-ориентированных информационных систем

Сергей Крывый *– Киевский национальный университет им. Т. Шевченко, Украина Киев-187, ГСП, 03680, просп. акад. Глушкова, 40; email: krivoi@i.com.ua*

Область исследований: Дискретная математика, теория автоматов, прикладная математическая логика, верификация программного обеспечения, программирование с ограничениями, онтологии.

Николай Петренко *– Ин-т кибернетики им. В.М. Глушкова НАН Украины, Киев-187 ГСП, 03680, просп. акад. Глушкова, 40; e-mail: petrng@ukr.net*

Область исследований: Методология и инструментальные средства автоматизированного проектирования онтологий предметных областей, системная интеграция междисциплинарных научных знаний

Дмитрий Бибиков *– Ин-т кибернетики им. В.М. Глушкова НАН Украины, Киев-187 ГСП, 03680, просп. акад. Глушкова, 40; e-mail: gbbcoff@gmail.com*

Область исследований: Искусственный интеллект, автоматизация поиска доказательств в формальных логических языках