АРХИТЕКТУРНО-СТРУКТУРНЫЕ ОСОБЕННОСТИ СРЕДСТВ АВТОМАТИЗАЦИИ ПРОЦЕССА ИЗВЛЕЧЕНИЯ ЗНАНИЙ ИЗ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТОВ

Александр Палагин, Сергей Крывый, Николай Петренко

***Abstract**: Рассматриваются проблемы архитектурно-структурной организации системы анализа естественно-языковых текстов с целью извлечения знаний. Предлагается автоматизированный итеративный подход к реализации такого анализа и некоторые архитектурно-структурные решения.*

***Keywords**: автоматизация обработки ЕЯТ, извлечение знаний, архитектура системы.*

***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 человеку необходимы средства автоматизированного поиска и обработки этой информации. В связи с этим возникает необходимость в разработке соответствующих средств автоматизации процессов поиска и обработки информациис целью извлечения из них релевантной запросу пользователя информации. Создание таких средств наталкивается на сложность проблемы анализа естественно языковых текстов, которая объясняется сложностью и неоднозначностью такого объекта, каким является естественный язык. Кроме семантической многозначности предложений естественного языка существует еще ряд проблем, связанных с анализом. Основными из них являются проблемы определения тематики, к которой принадлежит данный текст, смыслового содержания текста, эмоциональной окраски фраз, анализа фраз иронического и иносказательного характера, анализав условиях неполной или недостоверной информации (подразумевается нечто по умолчанию или вообще неизвестно) и т. п. Известно, что такого типа проблемы плохо формализуются,чем и объясняются причины сложности проведения такого анализа.

Данная работа является продолжением исследований авторов в этой области, начатых в [5-11].

Необходимые сведения

Как было сказано выше, тематика данной статьи относится к области искусственного интеллекта. Под понятием «искусственный интеллект» понимается«*некоторое устройство (созданное человеком) представляет собой **искусственный интеллект**, если, ведя с ним диалог по достаточно широкому кругу вопросов, человек не сможет различить, разговаривает он с разумным живым существом (например, с другим человеком) или с автоматическим устройством»* [1].При этом предполагается, что никаких дополнительных органов чувств (зрения, слуха, образов, мимики, жестов) искусственному интеллекту не нужно, а необходимы только те, с помощью которых он может вести диалог. Это означает,

что участники диалога должны иметь возможность обмениваться сообщениями (не видя друг друга) в той или иной (фиксированной) знаковой системе. Следовательно, система искусственного интеллекта имеет средства, которые способны воспринимать и формировать подобные сообщения, обладать памятью, которая дает возможность запоминать и хранить необходимую информацию, как вкладываемую в него заранее, так и получаемую в процессе диалога.

В случае анализа текста с целью извлечения из него знаний ситуация выглядит таким образом. Каким-то образом исходная информация о содержимом текста заносится в память, а затем в процессе диалога эта информация обрабатывается. При этом возникают две задачи:

1. получение исходной информации с занесением ее в память системы искусственного интеллекта;
2. обработка исходной информации с целью определения возможных противоречий и их устранения, или определения причин их возникновения;

Обе задачи сложны сами по себе, однако в данное время уже накоплены некоторые средства, с помощью которых можно частично решать обе задачи (о них речь пойдет дальше). При этом возникает третья задача:

3. в процессе диалога и анализа имеющейся информации необходимо учитывать появившийся опыт и новые знания (как следствия этого процесса), что приводит к необходимости модернизации, как знанийсистемы искусственного интеллекта, так и знаний человека, ведущего с ней диалог.

Характеристика типов знаний

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

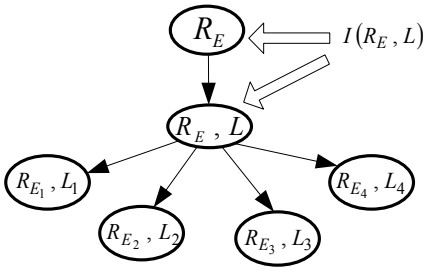
$$R_E = \left\{ R_{p_i} : p_i \in P \right\},$$

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[15–17].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. $L = L_1 \cup ... \cup L_j \cup ... \cup L_k$ – 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). Этот граф является основой построения онтологии той предметной области (ПдО), к которой относится анализируемый текст. 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 R_S. 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 конкретизации 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 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 рассматриваются предложения естественного языка, истинность того или иного факта зависит от индивидуума, анализирующего этот текст.

Мы принимаем 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 к понятиям, фигурирующим 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.

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» [4], 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 (разное понимание одного и того же предложения является причиной 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 [3].

Отметим также, что многие исследователи этой проблемы также считают, что моделирование знаний, извлеченных из ЕЯТ, не может ограничиваться формализацией только лишь непогрешимого интеллекта. 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.

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[2]. 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 целях более точной формулировки понятий «знание» и «извлечение знаний», которыми будем пользоваться в этой работе, рассмотрим следующие определения, пользуясь нотацией констрейнтного программирования [12].

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 (это отношение представлено в онтографе вершинами синонимичных объектов, которые факторизуются по предметно-ориентированному отношению синонимии). Переменные из

[2] Один из вариантов таких определений мы привели в [10].

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, а также дает возможность предложить следующий итеративный метод анализа.

Автоматизированный итеративный метод анализа предложений ЕЯ

Исходя из выше сказанного, можно предложить такой итеративный способ автоматизированной обработки ЕЯТ [11].

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, выделениеотглагольных 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 [13.14.]

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. Выполнитьанализ 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 [5, 15–17].

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(пользователями), являющимся автором 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. 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 [4], 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.

Предлагаемый итеративный метод реализуется с помощью онтолого-управляемой информационной системы (ОУИС), архитектурно-структурные особенности которой описываются ниже.

Особенности архитектурно-структурной организации ОУИС

Приведенный выше итеративный алгоритм обработки знаний, полученных из естественно-языкового текста, накладывает свои ограничения и требования на его архитектурно-структурную реализацию. Рассмотрим кратко свойства онтолого-управляемой архитектуры и онтологической компоненты знание-ориентированной информационной системы (ЗОИС) обработки предметно-ориентированных знаний в произвольной предметной области. По онтологической классификации его (базис) можно разделить на онтологическую подсистему и механизм онтологического управления. Они, в своем единстве, являются ключевой компонентой ЗОИС с естественно-языковым представлением, обработкой и актуализацией знаний, которой присущи интегрированные, взаимосвязанные и знание-ориентированные свойства и процедуры работы (в широком смысле) со знаниями и (само) развивающейся системы.

Модель обобщенной архитектуры развивающейся знание-ориентированной информационной системы описывается четверкой:

$$A = \langle O, D, F, P \rangle,$$

где $O = \{O_1, O_2, ..., O_k\}$ – множество онтологий, входящих в онтологическую подсистему ЗОИС;

$D = \{D_1, D_2, ..., D_k\}$ – множество онтологических знаний и данных, входящих в онтологии $\{O_1, O_2, ..., O_k\}$ соответственно;

$F = F^O \cup F^D$, где $F^D = \{F_1^D, F_2^D, ..., F_m^D\}$, $F^O = \{F_1^O, F_2^O, ..., F_l^O\}$ – множество процедур (задач), присущих для (само) развивающейся ЗОИС, в том числе, и задач пользователей, реализующих процессы на онтологиях и данных;

$P = \{P_O\} \bigcup \{P_D\}$, где $P_O = \{P_{O_1}, P_{O_2}, ..., P_{O_n}\}$ – предикаты, определенные на онтологиях $\{O_1, O_2, ..., O_k\}$, а $P_D = \{P_{D_1}, P_{D_2}, ..., P_{D_k}\}$ – предикаты, определенные на данных $\{D_1, D_2, ..., D_k\}$.

На следующем (более низком) уровне получим две многоосновные алгебраические системы:

$$AC_O = \left(\{O_1, O_2, ..., O_k\},\ \Omega_O = \left(\{F_1^O, F_2^O, ..., F_l^O\}, P_O\right)\right)$$

$$AC_D = \left(\{D_1, D_2, ..., D_k\},\ \Omega_D = \left(\{F_1^D, F_2^D, ..., F_m^D\}, P_D\right)\right).$$

AC_O и AC_D – алгебраические системы, представляющие онтологии и данные, относящиеся к этим онтологиям соответственно.

Пользователь, взаимодействуя с ЗОИС, формирует процедуры, реализующие соответствующие процессы верхнего уровня различного назначения с помощью операторов суперпозиции и итерации из сигнатур P_O и P_D, а также F_i^O и F_j^D, $i = \overline{1,l}$, $j = \overline{1,m}$, а на нижнем уровне система интерпретирует операторы и операции алгебраических систем AC_O и AC_D.

Выполненная формализация позволяет синтезировать блок-схему ЗОИС с онтолого-управляемой архитектурой (рис. 1). Она включает знание-ориентированную подсистему и интерфейс пользователя, подсистему манипулирования (экстра) лингвистической информацией, онтологическую подсистему и подсистему базового процессинга. Сюда следует отнести и внешние источники информации как важную компоненту извлечения и пополнения знаний и данных в соответствии с целенаправленной деятельностью ЗОИС.

На рис. 1 к компонентам, входящим в состав ОУИС, добавлена подсистема базового процессинга. Она включает блоки: хранилище данных и знаний; машину вывода базового процессинга; процедуры отработки целевых заданий и методы (алгоритмы) решения задач (вместе с интерпретатором базовых предикатов из P_O, P_D и процедур F_i^O и F_j^D). Первый блок включает ряд библиотек и баз данных, в которых хранится, обновляется и актуализируется концептуальная информация и фактографические данные, а именно:

- *онтологии* домена прикладных областей;
- *функции интерпретации* или определения концептов онтологии, записанные на некотором ограниченном ЕЯ или в виде логических формул подходящей формальной теории;
- *ограничения* для фактографических баз данных, общие принципы и/или аксиомы, тождественно истинные для элементов-концептов;

- *значения по умолчанию*: информация, являющаяся предпочтительной для элементов (фактов и правил вывода) концептов;
- *поведение*: правила, управляющие сценариями манипулирования для каждого концепта и взаимодействия наборов концептов в случае реализации (само) развивающейся ИС;
- *модуль библиотеки справочной информации*, включающей электронные энциклопедии доменов и толковые словари предметных областей.

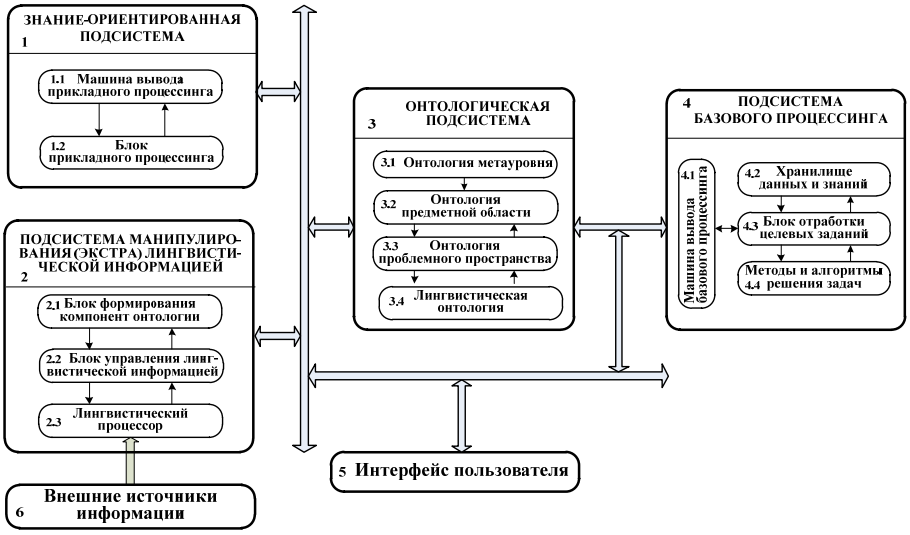


Рис. 1. Архитектура ЗОИС

Данный блок, совместно с блоком “Внешние источники информации” реализуют функции подсистемы “Информационный ресурс” инструментального комплекса онтологического назначения.

Остальные блоки реализуют функции, близкие к функциям *Решателя задач*.

Следует отметить, что в подсистему манипулирования (экстра) лингвистической информацией может быть включена знание-ориентированная поисковая система, выполняющая функции интеллектуального программного агента и являющаяся связующим звеном между ЗОИС и глобальным пространством Интернет. Кроме того, она организует хранение документов в виде онтологий текстовых документов в хранилище данных и знаний.

Такая композиция подсистем и блоков ЗОИС позволяет реализовать (в том числе) цепочку обобщенных процедур: “*обработка естественно-языковой информации → формальное логико-онтологическоепредставлениеЕЯ-информации → компьютерная обработка знаний*”, которые в известной степени соответствуют интегрированной информационной технологии работы со знаниями.

ОУИС функционирует в двух режимах:

1) отработка целевых заданий (внешних и внутренних, которые условно можно разделить на задания прикладного и базового процессинга соответственно), в частности, активизация процесса, актуализация информации, релевантной одной или нескольким ПдО, и размещение ее в памяти, решение поставленной проблемы (задачи), выработка, систематизация и выдача результирующих продукций (в случае знание-ориентированной деятельности – приращение знаний);
2) развитие ОУИС как информационной системы согласно общей стратегии развития: инвентаризация и систематизация знаний (расширение метазнаний), формализация и

когнитивизация представлений, интерпретационное расширение системы знаний, увеличение объема реакций и ассоциативных связей.

Основными функциями ОУИС являются:

- эффективное компактное представление системы знаний конкретной ПдО на базе современных ИТ (спецификация, концептуализация);
- поиск информации в системе знаний ПдО (справочные, обучающие системы);
- поиск необходимой информации в пространстве Интернет;
- постановка и решение прикладных задач в заданной ПдО (научных исследований и экспериментов, проектирования объектов новой техники, технологий и др.);
- развитие системы и получение новых знаний в соответствии с концептуальной моделью обработки знаний;
- диалог с пользователем.

Остальные функции ОУИС уточняются в зависимости от предметной области и задач, которые должны решаться в этой предметной области.

Заключение

Описанные в данной работе понятия и подходы к автоматизации обработки ЕЯТ составляют основу как теоретического, так практического решения проблемы 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 онтолого-управляемого алгебро-логического подхода к представлению и обработке текстовой 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 разных естественных языках.

Литература

1. Тьюринг А. Может ли машина мыслить? М.: Мир. – 1960.

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

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

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

5. 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.

6. Палагин А.В., Крывый С.Л., Бибиков Д.С. Обработка предложений естественного языка с использованием словарейичастотыпоявлениялов. – Natural and Artificial Intelligence Intern. Book Series. – Intelligent Processing. – ITHEA. – Sofia. – N 9. – 2010. – P. 44–52.

7. Алгебро-логічний підхід до аналізу та обробки текстової інформації / [Палагін О.В., Кривий С.Л., Петренко М.Г., Бібіков Д.С.]. – Проблеми програмування. Спеціальний випуск. – 7-а міжн. наук.-практ. конф. з програмування "УкрПРОГ'2010". – [Україна, Київ], 25-27 травня, 2010 р. – № 2, 3. – С. 318–329.

8. 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. – Проблеми програмування. Спеціальний випуск.

– 7-а міжн. наук.-практ. конф. з програмування “УкрПРОГ’2010”. – [Україна, Київ], 25-27 травня, 2010 р. – № 2, 3. – С. 382–389.

9. Крывый С.Л. Бибиков Д.С. Итеративный подход к анализу естественно-языковых текстов: логический аспект. – Проблеми програмування. Спеціальний випуск. – 8-а міжн. наук.-практ. конф. з програмування “УкрПРОГ’2012”. – [Україна, Київ], 25-27 травня, 2012 р. – № 2–3. – С. 10–17.
10. Палагин А.В. Онтологические методы и средства обработки предметных знаний / А.В. Палагин, С.Л. Крывый, Н.Г. Петренко. – [монография] – Луганск: изд-во ВНУим. В. Даля, 2012. – 324 с.
11. ПалагинА.В., КрывыйС.Л., ПетренкоН.Г.Обавтоматизациипроцессаизвлеченияязнанийизестественно-языко-выхтекстов. – Natural and Artificial Intelligence Intern. Book Series. – Intelligent Processing. – ITHEA. – Sofia. – N 9. – 2012. – P. 44–52.
12. Cohen D., Jeavons P. The Complexity of Constraint Languages. In “Handbookof Constraint Programming. – EditedbyF. Rossi, P. vanBeekandT. Walsh. – 2006. – P. 245 – 280.
13. Тейз А., Грибомон П., Луи Ж. и др. Логический подход к искусственному интеллекту. Отклассической логики к логическому программированию. – М.: Мир. – 1990. – 429 с.
14. Тейз А., Грибомон П., Юлен Г и др. Логический подход к искусственному интеллекту. Отмодальной логики к логике баз данных. – М.: Мир. – 1998. – 494 с.
15. Леонтьева Н.Н., Семенова С.Ю. Семантический словарь РУСЛАН как инструментарийкомпьютерного понимания. – М.: МГГИИ. – 2003. – С. 41–46.
16. Леонтьева Н.Н. К теории автоматического понимания естественных текстов. Часть 1.Моделирование системы"мягкого понимания" текста: информационно-лингвистическая модель. – М., МГУ, 2000. – 43 с.
17. 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

АлександрПалагин *– 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*

Областьисследований:дискретная математика, теория автоматов, прикладная математическая логика, верификация программного обеспечения, программирование с ограничениями, онтологии.

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

Областьисследований: архитектура и структура знание-ориентированных информационных систем, методология и инструментальные средства автоматизированного проектирования онтологий предметных областей, аппаратные лингвистические процессоры, междисциплинарные и трансдисциплинарные научные исследования