УДК 004.8

О. Г. Махасоева, **Д. Е. Пальчунов**

Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия

Институт математики им. С. Л. Соболева СО РАН
пр. Акад. Коптюга, 4, Новосибирск, 630090, Россия

E-mail: olga.makhasoeva@gmail.com; palch@math.nsc.ru

АВТОМАТИЗИРОВАННЫЕ МЕТОДЫ ПОСТРОЕНИЯ АТОМАРНОЙ ДИАГРАММЫ МОДЕЛИ ПО ТЕКСТУ ЕСТЕСТВЕННОГО ЯЗЫКА *

Разрабатывается теоретико-модельный подход к извлечению знаний из текстов естественного языка. Подход основан на формальном представлении извлекаемых знаний в виде конечных подмножеств атомарных диаграмм алгебраических систем. Описаны и реализованы в виде программной системы методы автоматизированного построения атомарных диаграмм моделей по текстам на русском языке. Разработаны словари существительных-номинализаций и валентностей глаголов.

Ключевые слова: извлечение знаний, представление знаний, теоретико-модельные методы, анализ текстов естественного языка, алгебраическая система, модель, атомарная диаграмма.

Введение

Статья посвящена проблеме извлечения знаний из текстов естественного языка и формального представления извлеченных знаний. Для формализации знаний, извлеченных из текстов естественного языка, используется теоретико-модельный подход. Знания представляются в виде предложений логики предикатов первого порядка сигнатуры онтологии рассматриваемой предметной области и верхнеуровневой онтологии естественного языка. В качестве естественного языка рассматривается русский язык.

В настоящее время важность разработки онтологий предметных областей трудно переоценить [1]. Онтологии лежат в основе проекта «Семантическая паутина» (Semantic Web) [2–5]. Одним из источников онтологической информации – информации о смысле терминов, ключевых понятий предметной области – являются тексты естественного языка: научные статьи, обзоры, монографии, энциклопедии и энциклопедические словари. В таких текстах специалисты, эксперты в данной предметной области излагают современное толкование смысла понятий, на языке которых описывается рассматриваемая предметная область.

В данной работе мы используем разработанный ранее теоретико-модельный подход к разработке онтологий предметных областей [6–8].

* Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 14-07-00903_а и Сибирского отделения РАН, проект № 3 «Принципы построения онтологии на основе концептуализаций средствами логических дескриптивных языков».

Махасоева О. Г., Пальчунов Д. Е. Автоматизированные методы построения атомарной диаграммы модели по тексту естественного языка // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2014. Т. 12, вып. 2. С. 64–73.

ISSN 1818-7900. Вестник НГУ. Серия: Информационные технологии. 2014. Том 12, выпуск 2

Теоретико-модельные методы представления знаний

Теория моделей является универсальным и наиболее разработанным инструментом формального представления знаний о предметных областях. Ранее нами был сформулирован тезис, названный *тезисом Мальцева – Тарского*: «Всякое описание ситуации, которое с точки зрения человека является полным, точным и формальным, может быть представлено в виде алгебраической системы» [6]. Был предложен теоретико-модельный подход к формальному представлению онтологий предметных областей [6–8]. Знания о предметной области, в том числе и онтологические знания, можно представлять как при помощи алгебраической системы некоторой сигнатуры, так и в виде множества предложений логики предикатов первого порядка этой сигнатуры.

Представление онтологии предметной области в виде множества предложений логики предикатов реализуется в проекте «Семантическая паутина» (Semantic Web) [2]. В рамках этого проекта онтология описывается на языке OWL (точнее, на его диалекте OWL-DL), который может быть транслирован в логику описаний (Description Logic, DL). Таким образом, онтологические знания о предметной области представляются в виде множества предложений языка логики описаний.

В рамках теоретико-модельного подхода к представлению знаний можно выделить два способа формализации: семантический и синтаксический. Семантические методы основаны на использовании алгебраических систем. Частным случаем алгебраических систем являются модели – алгебраические системы, сигнатура которых состоит только из символов предикатов и констант. Синтаксические методы основаны на использовании теорий, в частности элементарных теорий алгебраических систем. И тот и другой подход обладает своими преимуществами и недостатками.

В настоящей работе мы осуществляем синтез семантического и синтаксического методов. Для формализации знаний, извлекаемых из текстов естественного языка, мы используем конечные фрагменты (подмножества) атомарных диаграмм моделей. Атомарной диаграммой модели называется множество истинных на ней атомарных предложений – предикатов от констант и равенств констант; при этом сигнатура модели пополняется дополнительными константами именами для каждого элемента.

Заметим, что своей атомарной диаграммой модель определяется с точностью до изоморфизма. Поэтому, с одной стороны, совокупность конечных фрагментов атомарных диаграмм модели полностью задает эту модель, т. е. полностью определяет семантику. Сдругой стороны, как сама атомарная диаграмма модели, так и ее фрагменты являются некоторыми множествами предложений, т. е. описание знаний при помощи фрагментов атомарных диаграмм лежит в русле синтаксического подхода. Таким образом, представление знаний при помощи фрагментов атомарных диаграмм моделей является синтезом семантического и синтаксического подходов.

Введем необходимые определения и обозначения. Сведения по теории моделей можно найти в [9; 10].

Мы рассматриваем модели $\mathfrak{A} = < A;\ \sigma > = < A;\ P_1,\ \dots,\ P_n,\ c_1, \dots,\ c_l >$ сигнатуры $\sigma = < P_1, \dots, P_n, c_1, \dots, c_l >$, где A – основное множество (универсум) модели, $P_1, \dots, P_n$ – символы предикатов, а $c_1, \dots, c_l$ – символы констант. Через $S(\sigma)$ обозначим множество предложений, т. е. формул без свободных переменных, сигнатуры σ. Предложение φ называется ∀-предложением, если $\varphi = \forall x_1 \dots \forall x_n \psi(x_1 \dots x_n)$, где ψ – бесквантворная формула. $\mathfrak{A} \vDash \varphi$ означает, что на модели $\mathfrak{A}$ истинно предложение φ. Множество $Th(\mathfrak{A}) = \{\varphi \in S(\sigma) \mid \mathfrak{A} \vDash \varphi\}$ называется элементарной теорией модели $\mathfrak{A}$.

Для модели $\mathfrak{A}$ сигнатуры σ обозначим $\sigma_A = \sigma \cup \{c_a \mid a \in A\}$; при этом считаем, что $c_a \notin \sigma$ при $a \in A$. Через $\mathfrak{A}_A$ обозначим модель сигнатуры σ_A, обеднение которой до сигнатуры σ совпадает с $\mathfrak{A}$ и $c_a^{\mathfrak{A}_A} = a$ при $a \in \mathfrak{A}$.

Обозначим:

$FD(\mathfrak{A}) = \{\varphi \epsilon S(\sigma_A) \mid \mathfrak{A}_A \vDash \varphi\} = Th\ \mathfrak{A}_A$ – полная диаграмма модели $\mathfrak{A}$.

$D(\mathfrak{A}) = \{\varphi \epsilon S(\sigma_A) \mid \mathfrak{A}_A \vDash \varphi$ и предложение φ – бесквантворное$\}$ – элементарная диаграмма модели $\mathfrak{A}$.

Пусть сигнатура σ состоит только из символов предикатов и констант и $\mathfrak{A}$ – модель сигнатуры σ. Предложение φ назовем атомарным, если

$$\varphi = (c_1 = c_2),\ \ \varphi = \neg(c_1 = c_2),\ \ \varphi = P(c_1, \dots, c_n)\ \ \text{или}$$
$$\varphi = \neg P(c_1, \dots, c_n),\ \ \text{где}\ \ P, c_1, \dots, c_n \in \sigma_A.$$

Атомарной диаграммой модели $\mathfrak{A}$ назовем множество предложений

$AD(\mathfrak{A}) = \{\ \varphi \epsilon S(\sigma_A) \mid \mathfrak{A}_A \vDash \varphi$ и предложение φ – атомарное$\}$.

Замечание.

1. $AD(\mathfrak{A}) \subseteq D(\mathfrak{A}) \subseteq FD(\mathfrak{A})$.

2. Атомарная диаграмма $AD(\mathfrak{A})$ модели $\mathfrak{A}$ аксиоматизирует ее элементарную диаграмму $D(\mathfrak{A})$, а именно: $D(\mathfrak{A}) = \{\varphi \in S(\sigma_A) \mid \varphi -$ бесквантворное предложение и $AD(\mathfrak{A}) \vdash \varphi\}$.

Заметим, что атомарная диаграмма $AD(\mathfrak{A})$ определяет модель $\mathfrak{A}$ с точностью до изоморфизма. В частности, атомарная диаграмма $AD(\mathfrak{A})$ однозначно определяет элементарную диаграмму $D(\mathfrak{A})$ и полную диаграмму $FD(\mathfrak{A})$. Поэтому задачу построения модели можно свести к задаче построения ее атомарной диаграммы.

Таким образом, наша задача – по тексту естественного языка строить фрагмент атомарной диаграммы модели, который будет формальным представлением информации, содержащейся в тексте. Мы строим фрагменты атомарных диаграмм для каждого предложения, а затем объединяем полученные множества атомарных предложений в единый фрагмент атомарной диаграммы модели.

При этом понятия, содержащиеся в исходном тексте естественного языка, являются не просто сигнатурными символами, они имеют определенный смысл. Этот смысл понятий специфицируется онтологиями: верхнеуровневой онтологией всего естественного (в данном случае русского) языка и онтологией определенной предметной области.

В рамках развиваемого нами теоретико-модельного подхода к формализации знаний мы используем следующее определение онтологии.

Определение. Онтологией предметной области SD назовем пару $O = < SA, \sigma >$, где σ – множество ключевых понятий предметной области и SA – множество аналитических предложений, описывающих смысл этих ключевых понятий.

Аналитические предложения – это предложения, истинность которых определяется только значениями входящих в них терминов (понятий) [6; 11].

С одной стороны, при помощи онтологий мы можем пополнять фрагмент атомарной диаграммы модели, которую строим. Для этого, например, можно использовать представленные в онтологии отношения между понятиями: синонимию, «общее – частное» и др.

С другой стороны, поскольку понятия, содержащиеся в тексте, имеют определенный смысл, описываемый онтологией, на модели, фрагменты атомарной диаграммы которой мы строим, должна выполняться онтология рассматриваемой предметной области, т. е. должно быть истинно множество аналитических предложений SA.

Здесь возникает проблема: как, имея только конечные фрагменты атомарной диаграммы модели, гарантировать истинность множества предложений SA на всей модели?

Заметим, что большинство разработанных на настоящий момент времени онтологий могут быть представлены как множества ∀-предложений. Действительно, такие наиболее популярные онтологические отношения между понятиями, как «общее – частное» и синонимия формулируются в виде ∀-предложений.

Решение указанной выше проблемы для случая ∀-онтологий, т. е., онтологий, содержащих только ∀-предложения, дает следующее

Предложение. Пусть сигнатура σ содержит только символы предикатов и констант, $\Gamma \subseteq S(\sigma)$, каждое $\varphi \in \Gamma$ является ∀-предложением и $\mathfrak{A} \in K(\sigma)$. Тогда:

а) $\mathfrak{A} \vDash \Gamma$ тогда и только тогда, когда для любого $\Delta \subseteq AD(\mathfrak{A})$ выполнено $\Gamma, \Delta \nvdash$;

б) $\mathfrak{A} \vDash \Gamma$ тогда и только тогда, когда для любого конечного $\Delta \subseteq AD(\mathfrak{A})$ выполнено $\Gamma, \Delta \nvdash$;

в) $\mathfrak{A} \vDash \Gamma$ тогда и только тогда, когда для любого конечного $\Delta \subseteq AD(\mathfrak{A})$ и любого предложения $\varphi \in \Gamma$ выполнено $\varphi, \Delta \nvdash$.

Доказательство. Докажем для случая сигнатуры, состоящей только из символов предикатов. Случай сигнатуры, состоящей из символов предикатов и констант, доказывается аналогично:

а) (⇒) пусть $\mathfrak{A} \models \Gamma$ и $\Delta \subseteq AD(\mathfrak{A})$. Тогда $\Gamma \subseteq FD(\mathfrak{A})$. Кроме того, $\Delta \subseteq AD(\mathfrak{A}) \subseteq FD(\mathfrak{A})$ и $\mathfrak{A} \models FD(\mathfrak{A})$. Значит, $FD(\mathfrak{A}) \nvdash$ и $\Gamma \cup \Delta \subseteq FD(\mathfrak{A})$, поэтому $\Gamma, \Delta \nvdash$;

в) (⇐) пусть $\mathfrak{A} \not\models \Gamma$, тогда найдется $\varphi \in \Gamma$ такое, что $\mathfrak{A} \not\models \varphi$, значит, $\mathfrak{A} \models \neg\varphi$. Предложение $\varphi = \forall x_1 \dots \forall x_n \psi(x_1 \dots x_n)$, где ψ – бесквантворная формула сигнатуры σ. Поскольку тогда $\mathfrak{A} \models \exists x_1 \dots \exists x_n \neg\psi(x_1 \dots x_n)$, найдутся элементы $a_1, \dots, a_n \in \mathfrak{A}$ такие, что $\mathfrak{A} \models \neg\psi(a_1 \dots a_n)$.

Рассмотрим множество $B = \{a_1 \dots a_n\}$. В силу того, что сигнатура σ состоит только из символов предикатов, подмножество $B \subseteq |\mathfrak{A}|$ модели $\mathfrak{A}$ определяет ее подмодель $\mathfrak{B} \subseteq \mathfrak{A}$ с $|\mathfrak{B}| = B$. Поскольку формула $\neg\psi(\bar{x})$ бесквантворная, $\mathfrak{A} \models \neg\psi(a_1, \dots, a_n)$, $a_1, \dots, a_n \in \mathfrak{B}$ и $\mathfrak{B} \subseteq \mathfrak{A}$, выполнено $\mathfrak{B} \models \neg\psi(a_1, \dots, a_n)$. Следовательно, $\mathfrak{B} \models \neg\psi(c_{a_1}, \dots, c_{a_n})$ и $\neg\psi(c_{a_1}, \dots, c_{a_n}) \in D(\mathfrak{B})$.

В силу замечания $AD(\mathfrak{B})$ аксиоматизирует $D(\mathfrak{B})$, поэтому $AD(\mathfrak{B}) \vdash \neg\psi(c_{a_1}, \dots, c_{a_n})$. Легко проверить, что $AD(\mathfrak{B}) \subseteq AD(\mathfrak{A})$.

Обозначим $\Delta = AD(\mathfrak{B})$. Тогда множество предложений Δ конечно, $\Delta \subseteq AD(\mathfrak{A})$ и $\Delta \vdash \neg\psi(c_{a_1}, \dots, c_{a_n})$. Следовательно, $\Delta \vdash \exists\bar{x}\neg\psi(\bar{x})$, поэтому $\Delta \vdash \neg\forall\bar{x}\psi(\bar{x})$. Стало быть, $\Delta, \forall\bar{x}\psi(\bar{x}) \vdash$.

Таким образом, мы нашли предложение $\varphi = \forall\bar{x}\psi(\bar{x}) \in \Gamma$ и конечное множество предложений $\Delta = D(\mathfrak{B}) \subseteq AD(\mathfrak{A})$ такие, что $\varphi, \Delta \vdash$ – противоречие с условием.

Из полученного противоречия следует, что $\mathfrak{A} \models \Gamma$.

Мы доказали (⇒) для пункта (а), из чего следует (⇒) для пункта (б) и (⇒) для пункта (в).

Кроме того, мы доказали (⇐) для пункта (в), из чего следует (⇐) для пункта (б) и (⇐) для пункта (а).

Предложение доказано.

В силу предложения, для того чтобы гарантировать истинность ∀-онтологии на модели, которую мы строим по тексту естественного языка, достаточно проверять, что любое предложение, входящее в эту онтологию, будет совместно с каждым конечным фрагментом атомарной диаграммы данной модели. Здесь следует отметить еще один очень важный момент. Поскольку конечный фрагмент атомарной диаграммы состоит из бесквантворных предложений, его совместность с ∀-предложением является алгоритмически разрешимой.

Этапы порождения атомарной диаграммы модели

Представим поэтапное описание процесса построения атомарной диаграммы модели.

На *начальном этапе* работы пользователь вводит текст на русском языке. Введенный текст анализируется с помощью стороннего приложения CognitiveDwarf (подробное описание продукта можно найти в [12]), которое выделяет из текста морфемы и синтаксические связи, нормирует используемые слова (например, для существительных приводит их к единственному числу именительного падежа). На этом начальный этап завершен – мы получили исходный материал для построения сигнатуры модели и множества атомарных предложений.

На *втором этапе* происходит построение сигнатуры модели, для этого используется информация о морфологическом составе текста, которую мы получили на первом этапе. Для каждого слова, в зависимости от части речи, порождается свой сигнатурный символ – константа или предикат. Например, прилагательные интерпретируются как предикаты, запись *интересная*(*книга*) означает, что конкретная книга *книга* является интересной. Глагол *Дать*([*кто*]? *x*, [*кому*?] *y*, [*что*?] *z*) – предикат с тремя именованными аргументами, т. е. трехместный предикат. *Дать*(*Вася*, *Петя*, *книга*) – такая запись может пониматься как «Вася дал Пете книгу».

Методы сопоставления сигнатурных символов частям речи опираются на теорию И. А. Мельчука «Смысл ↔ Текст». Заметим, что сигнатура по окончании работы второго этапа может быть отредактирована и дополнена пользователем либо достроена при обработке нового текста.

Третий этап – построение атомарных предложений, истинных на модели. Исходя из правил русского языка были составлены словари, на основе которых для предикатов модели определяется набор их аргументов. Например, аргументами предиката-глагола «*купить*» выступают вопросы *кто*? *что*? *за сколько*? *и др.*, так как существует возможность адекватно задавать перечисленные вопросы к глаголу *купить – кто купил*, *что купил*, и др.

В ходе исследования были разработаны алгоритмы автоматического заполнения аргументов предикатов так, чтобы связи, представленные в тексте, отображались на модели. *Троян* Hesperbot нанес удар – из этой фразы получаем предикат-глагол нанести с аргументами [кто/что?, что?, на что?], после обработки аргументы заполнятся следующим образом: нанести[кто? Hesperbot, что? удар, на что? X], где X – неизвестная константа, так как в тексте нет информации, отвечающей за аргумент «на что?».

Ввиду разнообразия и сложности русского языка пользователю дана возможность полностью или частично изменять полученные автоматически результаты.

На *четвертом этапе* строится атомарная диаграмма модели – совокупность атомарных предложений расширенной сигнатуры σ_A, истинных на модели. Полученные на предыдущем шаге атомарные предложения собираются в единую атомарную диаграмму либо совокупность нескольких атомарных диаграмм.

В памяти компьютера модель хранится в виде xml-файла. С помощью полученной программы возможна интеграция нескольких сохраненных атомарных диаграмм, соответствующих разным текстам естественного языка.

Алгоритмы определения сигнатуры модели. Для определения сигнатуры модели – извлечения множества ключевых понятий предметной области из текста естественного языка – проводится первичная обработка текста с помощью стороннего приложения CognitiveDwarf. Выходной файл этой программы содержит следующую служебную информацию для каждого слова из входного документа:

- часть речи;
- нормальная форма;
- падеж, род, число, время (если возможно).

Помимо морфологии выводится список синтаксических связей. Большинство из них описываются двумя связанными словами и типом связи (подлежащее – сказуемое, прямое дополнение и др.). Для каждой части речи были разработаны свои алгоритмы для интерпретации.

Глаголы, причастия и деепричастия. Каждому глаголу в тексте ставится в соответствие предикат сигнатуры модели. Такой предикат всегда имеет хотя бы два аргумента – константу-действие и объект действия. Объекты действия указывают на производителя действия, если таковой имеется.

Заметим, что элементами модели, которую мы строим по тексту естественного языка, могут быть не только объекты, но и *конкретные действия*. Конкретное действие – это действие, которое происходит здесь и сейчас и в которое в данный момент времени вовлечены конкретные объекты. Рассмотрение только моделей, т. е. алгебраических систем, сигнатура которых не содержит функциональных символов, избавляет нас от необходимости работать при этом с многосортными системами. Конкретные действия и модели представляются при помощи специальных констант: констант-действий.

Глаголы *шел*, *брел*, *наступал* в предложении «*шел я*, *брел я*, *наступал то с пятки*, *то с носка*» описывают одно и то же действие, хотя в строгом смысле синонимам не являются. Чтобы показать тождественность действий, используются одна и та же константа-действие: *шел*([*act*] ***act1***, ...), *брел*([*act*] ***act1***, ...), *наступал*([*act*] ***act1***, ...).

На рис. 1 показано, что пользователем было введено предложение «*вирусы меняют поведение программ*, *внедряют себя в их исполняемый код*», и результат ввода. Здесь описано одно действие – изменение программы посредством внедрения другого кода. Чтобы выразить это на языке моделей, мы использовали одну и ту же константу-действие «*менять_0*»:

- внедрять(*менять_0*, внедрять_obj, программа, код);
- менять(*менять_0*, вирус, поведение, на_что_менять_0).

Употребление такой константы овеществляет действие и может также быть использовано для разграничения действий во времени.

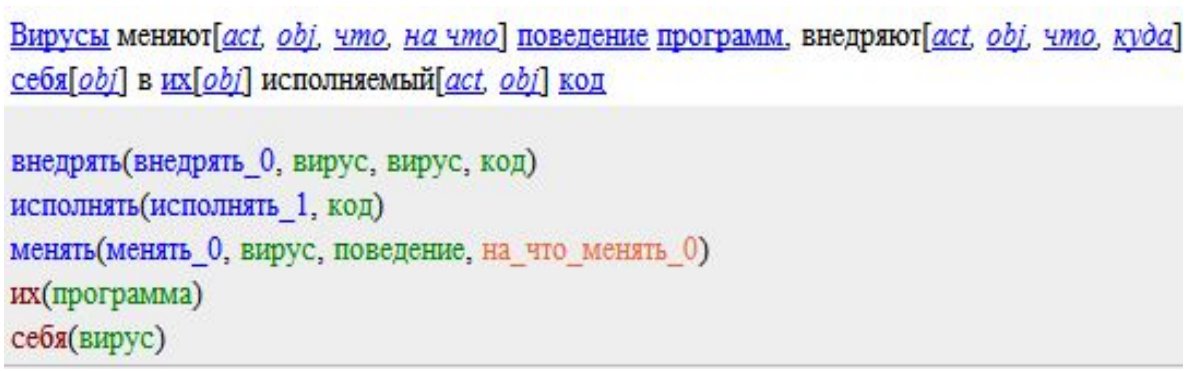

Рис. 1. Константа-действие «менять_0» в двух предикатах

Кроме двух основных аргументов глаголов – константы–действия и объекта действия, у предикатов могут быть дополнительные аргументы. Наличие дополнительных аргументов определяется правилами употребления слов русского языка. Принято различать не требующие дополнений глаголы (*смеркаться*, *греметь*) и глаголы, которые требуют дополнения каким-либо другим словом. К этому классу относятся, например, глаголы*: меняют*[*act*, *obj*, *что*, *на что*], *внедряют*[*act*, *obj*, *что*, *куда*].

Об этих аргументах, указывающих, что слова могут вступать в синтаксическую связь с другими словами (или требуют дополнения другими словами), принято говорить как о «валентностях» слов. Нами был создан словарь валентностей на основе словарей В. И. Даля, Т. Ф. Ефремовой, Д. Э. Розенталя. Он содержит более 2,3 тысяч слов и 75 различных типов вопросов (*что*, *за что*, *кем*, и т. д.).

Если вопрос невозможно задать к слову по правилам русского языка, порождается специальный запрет, который не позволяет заполнить некорректную валентность.

Причастия и деепричастия – это части речи, образованные от глаголов и обозначающие действия. Мы заменяем их на однокоренные глаголы, а затем добавляем предикаты в сигнатуру таким же образом, как это было проделано с глаголами. При такой замене не происходит искажения или потери смысла: фраза «*Вирусы*, ***меняющие*** *поведение программ*, ***внедряют*** *себя в их исполняемый код*» эквивалентна по смыслу фразе «*Вирусы* ***меняют*** *поведение программ*, ***внедряют*** *себя в их исполняемый код*», которая, в свою очередь, эквивалентна по смыслу фразе «*Вирусы* ***меняют*** *поведение программ*, ***внедряя*** *себя в их исполняемый код*».

Результат работы программы показан на рис. 2.

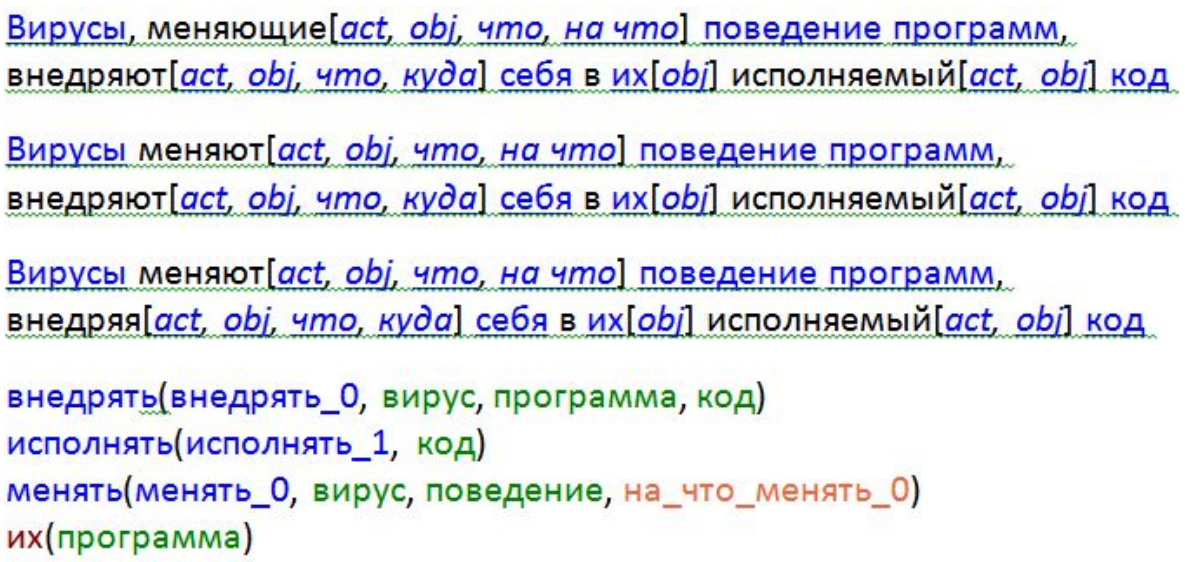

Рис. 2. Причастия и деепричастия

Прилагательные. Все прилагательные представляются как предикаты модели, арность (количество мест) которых равна единице. Они сопоставляют константу с описанной прилагательным характеристикой, показывая, обладает ли объект этой характеристикой. Например, при обработке словосочетания *интересная книга* получаем предикат сигнатуры модели, соответствующий прилагательному. Его единственное место (аргумент) – [*кто/что*] заполняется константой *книга*: *интересная*[*книга*]. Это означает, что книга обладает характеристикой интересности.

Существительные. В отличие от глаголов и прилагательных существительные могут быть представлены в сигнатуре как предикатами, так и константами.

В первую очередь все существительные подвергаются проверке на номинализацию. Номинализация – это «отглагольное» существительное, описывающее действие, а не объект. К таким относится *бег*, *удар*, *вынос* (от глаголов *бежать*, *ударять*, *выносить*) и др. Номинализации больше всего похожи на константы-действия, так как они не имеют под собой реального объекта, но отражают процесс исполнения определенного действия, например, *бег – бежать*. Номинализации представляются в сигнатуре как предикаты-глаголы: происходит поиск совпадений существительного-номинализации по словарю, в случае подтверждения извлекаются глаголы, от которых была образована номинализация. Используемый словарь был разработан на основе словарей А. А. Зализняка, Т. Ф. Ефремовой и содержит около 8 тысяч наименований.

На рис. 3 изображен результат работы программы по тексту «Международная антивирусная компания ESET сообщает об обнаружении новой модификации банковского трояна, которая обладает возможностями по краже биткоинов», выделено и заменено на глаголы 3 номинализации: *обнаружение*, *модификация*, *кража.*

Международная[*obj*] антивирусная[*obj*] компания ESET сообщает[*act, obj, о чем, что*] об обнаружении[*обнаружить, обнаружиться : act, obj, что*] новой[*obj*] модификации [*модифицировать : act, obj, что*] банковского[*obj*] трояна, которая[*obj*] обладает[*act, obj, чем*] возможностями по краже[*красть : act, obj, у кого, что*] биткоинов.

Рис. 3. Номинализации: обнаружение, модификация, кража

Если же существительное не является номинализацией, то в зависимости от того, конкретный объект имеется в виду (*Онегин*, *Гоголь*) или класс объектов (запись *кошка*[*X*] означает, что *X* принадлежит классу объектов *кошки*), существительному в сигнатуре модели ставится в соответствие либо константа, либо предикат.

На этапе редактирования модели пользователь имеет возможность изменить тип существительного – поменять константу на предикат и наоборот либо добавить существительное к словарю номинализаций.

Программная реализация

На данный момент времени с помощью разработанной нами программной системы пользователь может строить фрагмент атомарной диаграммы модели по тексту на русском языке (далее для удобства будем называть этот фрагмент моделью), редактировать и визуализировать полученную модель, сохранять ее, загружать снова одну из сохраненных моделей. При необходимости пользователь может создать и редактировать свои словари валентностей и номинализаций (например, для работы с текстом определенной предметной области). Программная система предоставляет возможность ответа на некоторые прямые вопросы, заданные к содержимому модели. Кроме того, если в тексте недостает информации для заполнения всех аргументов предиката, система задает пользователю наводящие вопросы. На рис. 4 отображена use-case-диаграмма приложения, демонстрирующая возможности пользователя. Большинство из перечисленных возможностей были проиллюстрированы ранее.

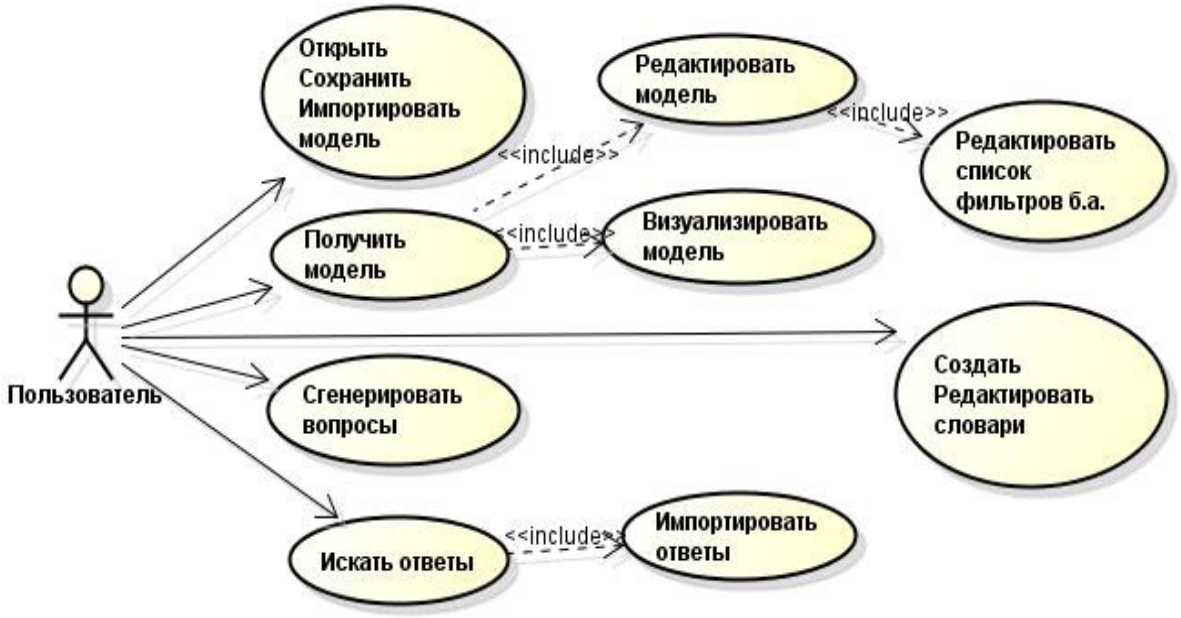

Рис 4. Use-case-диаграмма разработанной программной системы

На рис. 5 приведен пример работы программы.

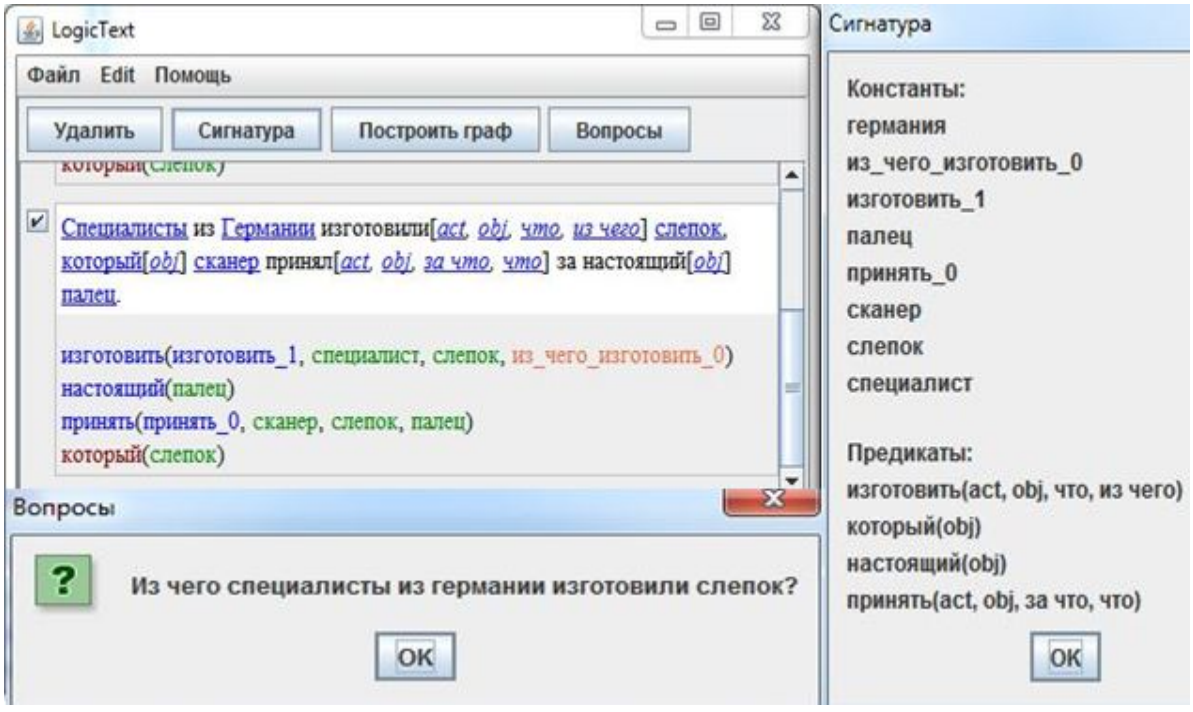

Рис. 5. Вопросы к недостающим валентностям (слева внизу), сигнатура (справа)

Граф, изображенный на рис. 6, визуализирует фрагмент атомарной диаграммы модели, что позволяет пользователю облегчить просмотр связей, присутствующих в модели. Визуализируется текст «*Ответ*, *который не содержит модель*, *ищите в сети*». Используется библиотека JUNG.

Обозначения: прямоугольники – предикаты (в том числе отрицания предикатов), овалы – константы (присутствующие в тексте непосредственно; опущенные константы; константы-действия). Связи изображаются именованными стрелками.

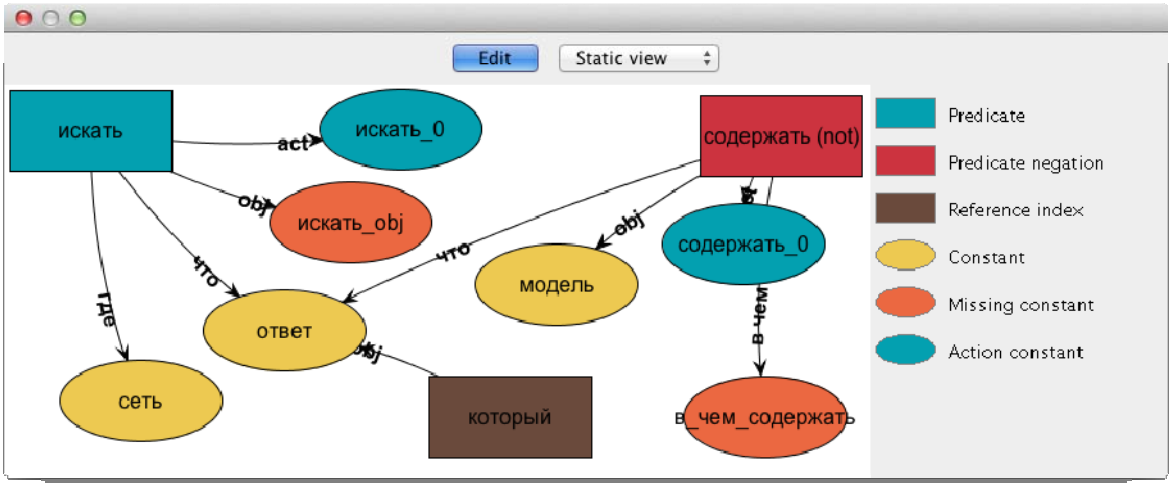

Рис. 6. Граф «Ответ, который не содержит модель, ищите в сети»

Заключение

В работе предложен теоретико-модельный подход к извлечению знаний из текстов естественного языка. В основе подхода лежит представление знаний при помощи конечных фрагментов атомарных диаграмм моделей. Разработаны методы интерпретации различных частей речи и синтаксических связей с целью автоматического порождения сигнатуры модели. На основе этого разработаны методы автоматического построения атомарных предложений данной сигнатуры по тексту естественного языка. В ходе исследования были созданы словари номинализаций (8 000 понятий) и валентностей (2 300 глаголов).

Разработана программная система, которая осуществляет порождение фрагмента атомарной диаграммы модели по тексту естественного языка. Программная система реализует разработанные методы и алгоритмы. Пользователь может редактировать и визуализировать автоматически построенный фрагмент атомарной диаграммы модели. Программная система предоставляет пользователю возможность получения ответов на вопросы определенного вида на основе знаний, представленных в модели.

Список литературы

1. *Staab S., Studer R.* (Eds.) The Handbook on Ontologies in Information Systems. Springer Verlag, 2003. 811 p.

2. *Daconta M. C., Obrst L. J., Smith K. T.* The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management. Wiley Publishing, 2003. 312 p.

3. *Fensel D.* OIL: An Ontology Infrastructure for the Semantic Web // IEEE Intelligent Systems. 2001. Vol. 16. P. 38–45.

4. *Maedche A.* Ontology Learning for the Semantic Web. Kluwer Academic Publishers, 2002. 244 p.

5. *McGuinness D., Harmelen F.* (Eds.) OWL Web Ontology Language Overview. URL: http://www.w3.org/TR/owl-features/

6. *Пальчунов Д. Е.* Моделирование мышления и формализация рефлексии. I: Теоретико-модельная формализация онтологии и рефлексии // Философия науки. 2006. № 4 (31). С. 86–114.

7. *Пальчунов Д. Е.* Моделирование мышления и формализация рефлексии. II: Онтологии и формализации понятий // Философия науки. 2008. № 2 (37). С. 62–99.

8. *Пальчунов Д. Е.* Решение задачи поиска информации на основе онтологий. // Бизнес-информатика. 2008. № 1. С. 3–13.

9. *Ершов Ю. Л., Палютин Е. А.* Математическая логика. М.: Наука, 1979. 317 с.

10. *Кейслер Г., Чэн Ч. Ч.* Теория моделей. М.: Мир, 1977. 615 с.

11. *Carnap R.* Meaning and Necessity. A Study in Semantics and Modal Logic. Chicago, 1956. 220 p.

12. *Антонова А. А.* Синтаксический анализатор для русского и английского языков // Сб. тр. ИСА РАН. Информационно-аналитические аспекты в задачах управления: М.: ЛКИ, 2007. Т. 29. С. 329–337.

Материал поступил в редколлегию 06.06.2014

O. G. Makhasoeva, D. E. Palchunov

SEMI-AUTOMATIC METHODS OF A CONSTRUCTION OF THE ATOMIC DIAGRAMS FROM NATURAL LANGUAGE TEXTS

The paper is devoted to a model-theoretical approach to the knowledge extraction from the natural language texts. The approach is based on a formal representation of extracted knowledge in terms of the finite subsets of atomic diagrams of algebraic systems. Methods of a semi-automatic construction of the atomic diagrams from texts in Russian are described in the paper and are implemented as a program system. A set of dictionaries (nominalizations and verbs valences) was developed.

Keywords: knowledge extraction, knowledge representation, model-theoretical methods, analysis of natural language texts, algebraic system, model, atomic diagram.

References

1. Staab S., Studer R. (Eds.) The Handbook on Ontologies in Information Systems. Springer Verlag, 2003, 811 p.

2. Daconta M. C., Obrst L. J., Smith K. T. The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management. Wiley Publishing, 2003, 312 p.

3. Fensel D. OIL: An Ontology Infrastructure for the Semantic Web. IEEE Intelligent Systems, 2001, vol. 16, p. 38–45.

4. Maedche A. Ontology Learning for the Semantic Web. Kluwer Academic Publishers, 2002, 244 p.

5. McGuinness D., Harmelen F. (Eds.) OWL Web Ontology Language Overview. URL: http://www.w3.org/TR/owl-features/.

6. Palchunov D. E. Modelirovanie myshleniya i formalizaciya refleksii. I: Teoretiko-model'naya formalizaciya ontologii i refleksii [Modeling of reasoning and formalization of reflection I: Model theoretical formalization of ontology and reflection]. *Filosofiya nauki*, 2006, no. 4 (31), p. 86–114. (In Russ.).

7. Palchunov D. E. Modelirovanie myshleniya i formalizaciya refleksii. II: Ontologii i formalizacii ponyatij [Modeling of reasoning and formalization of reflection. II: Ontologies and formalization of concepts]. *Filosofiya nauki*, 2008, no. 2 (37), p. 62–99. (In Russ.)

8. Palchunov D. E. Reshenie zadachi poiska informacii na osnove ontologij [The solution of the problem of information retrieval based on ontologies]. *Bisnes-informatika*, 2008, no. 1, p. 3–13. (In Russ.).

9. Ershov Yu. L., Palyutin E. A. Matematicheskaya logika [Mathematical Logic]. Moscow, Nauka, 1979, 317 p. (In Russ.)

10. Chang C. C., Keisler H. J. Model theory. Moscow, Mir, 1977, 615 p. (In Russ.).

11. Carnap R. Meaning and Necessity. A Study in Semantics and Modal Logic. Chicago, 1956, 220 p.

12. Antonova A. A. The development of a syntactic parser for Russian and English. Collection of scientific papers of ISA RAS. Information-analytical aspects in control problems. Moscow, LKI Publisher, 2007, vol. 29, p. 329–337. (In Russ.).