УДК 004.891.2

М.И. Кочергин, К.С. Кочергина

Формализация текстовых условий задач по физике

Рассматривается процедура автоматизации формализации текстовых описаний физических задач. Предлагается схема трехэтапной формализации текстовых описаний объектов и процессов различных предметных областей. Приведены алгоритмы формализации на текстовом, предметном и поведенческом уровнях. Процедура формализации иллюстрируется на примере задачи по физике из раздела кинематики.

Ключевые слова: формализация, физика, извлечение информации.

doi: 10.21293/1818-0442-2016-19-1-65-68

Формализация текстовых условий задачи по физике является необходимой составляющей процедуры её решения. Под формализацией будем понимать процесс выявления важных характеристик и параметров объекта и описание их в строгой заданной форме. Для любой предметной области можно выделить три уровня автоматизированной формализации текстовых описаний:

1) формализация описания на уровне текста – выделение из текстового описания сущностей, их признаков, действий, а также отношений с другими сущностями (получаемое при этом представление имеет вид семантического графа, где вершинами являются слова, а дугами – связывающие их семантические отношения, например: «часть–целое», «агент–реципиент» и др.);

2) формализация описания на предметном (аспектном) уровне – объективизация и параметризация полученного представления методом многоаспектного анализа [1] с использованием знаний из предметной области, в ходе которого оно получает вид отдельных взаимосвязанных наборов вида «объект–действие–параметр»;

3) формализация описания на поведенческом (модельном) уровне – структуризация полученного представления (выявление неявных и математическое описание найденных ранее межобъектных связей), в ходе которой оно приобретает более сложную единую форму. Так, задачи по физике отображаются с помощью диаграмм поведения, характеризующих связи объектов в задаче и смены их поведений. Вводимые на данном уровне модели поведения и отношений объектов описывают их действия, состояние и взаимосвязи с помощью конкретных физических и математических закономерностей (например, «Равномерное прямолинейное движение: $v = s \cdot t$ » или «Равноускоренное прямолинейное движение: $v = v_0 + a \cdot t$ »).

Задачи по физике как объект формализации

Одним из простых примеров описаний физико-химических процессов являются тексты задач по физике:

1) в них содержится описание некоторого физико-химического процесса (условие задачи);

2) в них ставится некоторая проблема, которую необходимо решить (требования задачи);

3) для их решения необходимо знание предметной области – физических законов.

Другими словами, решение задачи по физике можно рассматривать как процедуру исследования некоторого явления или процесса, требующего экспертных знаний в соответствующей предметной области (физика) и поиска подходящего решения поставленной проблемы (нахождение значения искомого параметра).

Разработка методов перевода (трансляции) текстовых описаний задач такого типа из естественно-языкового представления в формальное, пригодное для машинной обработки, позволяет автоматизировать процесс поиска их решения, что, в свою очередь, является отправной точкой для решения более сложных задач в различных предметных областях (например, в моделировании управляемых технических или эколого-экономических систем).

Перевод текстовых описаний на формальный язык

Для осуществления перевода текстовых описаний процессов и систем различных предметных областей (ПО) на формальный язык необходимо применение методов компьютерной лингвистики (автоматического анализа текста) и интеллектуального анализа данных (ИАД). Под ИАД понимается обнаружение интеллектуальной системой нового (относительно её базы фактов) знания [2]. В частности, к ИАД может быть отнесена и процедура автоматизированной формализации текстовых описаний объектов различной природы. Такими описаниями могут быть тексты физических задач (для ПО «Физика») или описания природоохранных мероприятий (для ПО «Эколого-экономические системы»).

Автоматизированная формализация физических задач состоит из двух шагов:

1) предобработка текстовых условий задачи – построение их предварительного формализованного представления (формализация на уровне текста), минимально иллюстрирующего содержание задачи с помощью семантических отношений между выделенными в ней сущностями;

2) непосредственно интеллектуальная обработка полученного графа – извлечение из него необходимой полезной информации и приведение самой задачи к формализованному виду (сначала на предметном уровне, затем на поведенческом).

На первом этапе применяются методы компьютерной лингвистики, производится автоматический анализ текстовых условий задачи, в результате которого строится семантический граф (такое представление может быть представлено и в другом виде, например, в предикатном). Минимальной (атомарной) единицей формализуемого представления на данном этапе является слово.

На втором этапе производится интеллектуальная обработка полученного представления, заключающаяся в его объективизации, параметризации и структуризации. Результат данного этапа – формальное представление (информационная модель) описываемого в тексте процесса или объекта (системы объектов), пригодное для дальнейшей постобработки. Минимальной единицей такого представления является уже не слово, а модель объекта или его фрагмента, выраженная на математическом языке и описывающая его. Так, задача по физике на дан-

ном этапе представляется в виде диаграммы, иллюстрирующей смену поведения (описываемого законами физики) действующими объектами и взаимных качественных и количественных связей этих объектов. Такое представление, на наш взгляд, удобно описывать терминами теории моделирования: с помощью объектных моделей (ОМ) – физических законов, характеризующих поведение или состояние объекта в определённый момент времени, и моделей отношений (МО) – физико-математических законов, характеризующих взаимосвязи параметров различных объектов.

Предобработка текстовых условий задачи

Предобработка текстовых описаний задач по физике осуществляется разработанным модулем формализации текстовых условий [3]. Содержание исполняемых процедур на шаге предобработки представлено на рис. 1.

Рис. 1. Схема процедуры предобработки текстовых условий задачи

Так, на первом этапе предобработки (графематический анализ) происходит простое разбиение текстового потока на минимальные единицы, разделённые пробелом, и их группировка (например, обозначение физической величины латинскими буквами группируется вместе с числовым значением и единицами измерения). Выходные данные этого этапа – множество T , представляющее последовательность текстовых единиц (ТЕ). ТЕ представляет собой абстрактную структуру данных с полями знаковой формы (слова в предложении) и набора числовых параметров.

На втором этапе (морфологический анализ) происходит классификация ТЕ (определение частеречной принадлежности слов к одному из классов множества C) на основании комбинированного (статистического и словарного) метода. Словарный метод используется для определения слов-исключений и условных обозначений (единицы измерения, латинские обозначения величин и т.д.), статистический – в остальных случаях. В соответствии с тем, к какому классу отнесён объект, ему присваиваются значения морфологических параметров (вектор M) из соответствующих составленных таблиц. Эти параметры несут информацию о характере рассматриваемого слова для определения его связи с остальными словами в предложении, а затем и в тексте задачи. Выходными данными этого этапа являются тройки (T, C, M) , устанавливающие гомоморфное отображение каждой ТЕ в некоторый класс (часть речи) и вектор числовых (морфологических) параметров.

На третьем этапе (синтаксический анализ) происходит анализ морфологических характеристик ТЕ для построения ориентированного графа G , отображающего их синтаксические связи S в предложении, что даёт первоначальное представление о семантике слов. Выходными данными этого этапа является граф $G(T, S)$ синтаксических связей ТЕ.

На четвертом этапе (поверхностный семантический анализ) происходит распределение семантических ролей R по словам с использованием данных, полученных с предыдущих этапов. Семантические роли – минимальные смысловые категории слов, более полно характеризующие их семантическое значение (чем синтаксические функции), подобранные в ходе исследования на основе семантических падежей Ч. Филмора [4]. Выходные данные этого этапа – взвешенный ориентированный семантико-синтаксический граф $G(T, S, R)$, где веса R синтаксического графа $G(T, S)$ – семантические роли ТЕ.

Этап предобработки итеративен для каждого предложения в тексте задачи. Для исключения ошибок классификации или взвешивания графов на каждом этапе выполняется проверка непротиворечивости данных с предыдущего этапа, при необходимости инициируется перезапуск предыдущего этапа с пересчётом результата.

Отметим, что для предобработки текстовых описаний также применимы и другие существующие программные средства (синтактико-семантические парсеры), но при этом требуется их некоторая адаптация и дополнительная обработка полученного ими результата для корректной параметризации и

структуризации формального представления, а также не гарантируется высокая точность предобработки.

Формализация текстовых условий физических задач на предметном уровне

На данном этапе производится объективизация и параметризация строящегося формализованного представления задачи по физике. Для этого производится упрощение сформированного синтактико-семантического графа и приведение его к структуре «объект–действие–параметр–значение». Указанная процедура осуществляется посредством поиска ключевых слов, занесённых в базу терминов предметной области, среди вершин графа. Фрагмент базы данных представлен в таблице.

Фрагмент базы ключевых слов в задачах физики

Слово	Роль	Категория
Ехать	Действие	Движение
Скорость	Параметр	Параметр действия
км/ч	Единица измерения	Скорость
Равномерно	Характер	Действие

После того как найдено ключевое слово, ему присваивается новая функциональная роль (вместо семантической) из соответствующего (одноимённого) столбца. Данные из столбца «категория» являются вспомогательными, и используются для проверки адекватности формализованного представления и интерпретируются в зависимости от роли ключевого слова. Так, для роли «действие» они сообщают о характере действия, для «параметра» – о принадлежности объекту или действию (например, «масса» – параметр объекта, а «пройденный путь» – действия), для «единицы измерения» – о параметре, который измеряется в этих единицах.

После того, как всем ключевым словам присвоены новые роли, производится построение нового графа с исключением вершин, не содержащих ключевых слов, кроме имеющих роль «агент» или «реципиент» или содержащих признаки объектов, необходимые для их дифференциации.

Формализация текстовых условий физических задач на поведенческом уровне

После формирования предметного представления задачи производится сопоставление каждому отдельному набору параметров физического закона

(объектной модели), выявляются математические зависимости параметров различных объектных поведений (состояний) друг от друга (модели отношений), а также производится окончательная структуризация задачи. Указанная процедура производится интерпретатором продукционных правил, хранимых в базе знаний предметной области. Интерпретатор считывает правило из базы знаний и последовательно проверяет выполняемость условий, записанных в антецеденте. В случае истинности антецедента запускается выполнение действий, записанных в консеквенте, в случае ложности выполняется проверка следующего правила. Дополнить базу правил может эксперт в некоторой предметной области, не обращаясь за помощью к техническому специалисту. Единственным требованием является соблюдение синтаксиса языка правил. Приведём пример такого продукционного правила: «ЕСЛИ (Параметры Действия СОДЕРЖИТ «скорость» & «расстояние» & «время») И (Параметры Действия НЕ СОДЕРЖИТ «ускорение» & «импульс» & «масса» & «кинетическая энергия» & «потенциальная энергия» & «начальное положение» & «текущее положение») ТО (Поведение = «Равномерное прямолинейное движение: $v=s/t$) КОНЕЦ ЕСЛИ;».

Формируемое на данном этапе представление является информационной моделью рассматриваемой физической задачи. Далее возможно применение средств постобработки, производящей, например, аналитические преобразования для вывода расчётной формулы или же формирования компьютерной модели задачи.

Пример формализации текстовых условий физической задачи

Рассмотрим простую физическую задачу из раздела кинематики: «Первую половину своего пути автомобиль двигался со скоростью 80 км/ч, а вторую – со скоростью 40 км/ч. Какова средняя скорость автомобиля на всем пути?» В данной задаче можно выделить один объект («автомобиль»), производящий смену поведения (изменение значения параметра «скорость» при равномерном прямолинейном движении). Проиллюстрируем поэтапно ход автоматической формализации рассматриваемой задачи. На рис. 2 изображен результат формализации задачи на уровне текста.

Рис. 2. Пример семантико-синтаксического графа задачи

Изображенный семантико-синтаксический граф поверхностно иллюстрирует смысловое содержание задачи только на уровне текста. Например, параметр «средняя скорость» во втором предложении приобретает роль «агента» (действующего лица), так как является подлежащим. Данное представление нуждается в дальнейшей обработке.

Далее производится объективизация и параметризация полученного графа, результат которой представлен на рис. 3.

Рис. 3. Формализованное представление задачи на предметном уровне

На данном уровне выделяется объект – «автомобиль», совершающий одно действие, обладающее тремя сменяющимися наборами параметров, разделённых связками (выраженными синтаксически на уровне языка), позволяющими определить порядок смены поведения объектом.

Затем производится сопоставление выделенных групп параметров с физическими законами, описанными в базе знаний предметной области, по критерию характерности переменной закону и окончательная структуризация задачи. Результат выполнения указанных процедур представлен графически на рис. 4.

Рис. 4. Формализованное представление задачи на поведенческом уровне

В данной задаче выделяются три поведения объекта (объектные модели: ОМ_1, ОМ_2, ОМ_3) и три связывающих их математические зависимости (модели отношений: МО_1, МО_2, МО_3). Полученное представление является окончательно формализованным и структурированным и пригодно для постобработки.

Заключение

Разработанная система производит формализацию текстовых условий задач по физике. Формализация производится в три этапа: на уровне текста (с применением методов компьютерной лингвистики), на предметном уровне (посредством объективизации и параметризации условий задачи с использованием базы ключевых слов предметной области) и поведенческом уровне (посредством анализа и интерпретации продукционных правил из базы знаний предметной области). Предлагаемая схема является универсальной для различных предметных областей и обуславливается содержанием базы знаний и базы терминов (ключевых слов). В настоящее время производится адаптация разработанных программных средств для формализации описаний эколого-экономических систем.

Литература

1. Дмитриев В.М. Метод многоаспектного анализа как алгоритм формализации задач по физике / В.М. Дмитриев, А.Ю. Филиппов, О.Н. Шарова // Вестник ТГПУ. – Томск: Томский гос. пед. ун-т, 2005. – № 4. – С. 60–66.
2. Финн В.К. Об интеллектуальном анализе данных // Новости искусственного интеллекта. – 2004. – № 3. – С. 3–19.
3. Кочергин М.И. Особенности модуля автоматического анализа в системе логико-лингвистического анализа текстов задач по физике / М.И. Кочергин, К.С. Спиридонова // Научная сессия ТУСУР–2014: матер. Всерос. науч.-техн. конф. студентов, аспирантов и молодых ученых, Томск, 14–16 мая 2014 г.: в 5 ч. – Томск: В-Спектр, 2014:– Ч. 2. – С. 257–260.
4. Filmor C. Frames and the semantics of understanding / C. Filmor // Quaderni di semantica. – Vol. 4/2. – Dec. – 1985. – P. 222–254.

Кочергин Максим Игоревич

Аспирант каф. моделирования и системного анализа (МиСА) ТУСУР
Тел.: 8 (382-2) 41-39-15
Эл. почта: max24kochergin@gmail.com

Кочергина Кристина Сергеевна

Магистрант 2-го года обучения филол. фак-та ТГУ
Тел.: 8 (382-2) 52-70-44
Эл. почта: tinakochergina@gmail.com

Kochergin M.I., Kochergina K.S.

Formalization of text descriptions for physics tasks

The paper presents an automation technique aimed to formalize text descriptions of physics tasks. Three-stage formalization scheme is suggested. Algorithms for the formalization on text, substantive and behavior levels are given. The example for the task formalization in the field of the kinematics is provided.

Keywords: formalization, physics, information extraction.