УДК 004.82:004.912

ПОДХОД К МОДЕЛИРОВАНИЮ ПРОЦЕССА ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ ИЗ ТЕКСТА НА ОСНОВЕ ОНТОЛОГИИ

Е.А. Сидорова

Институт систем информатики им. А.П. Ершова СО РАН, Новосибирск, Россия
lsidorova@iis.nsk.su

Аннотация

В статье рассматриваются модели и методы представления знаний, ориентированные на задачи автоматической обработки текста и извлечения информации. В рамках предлагаемого подхода извлечение информации рассматривается как процесс пополнения онтологии информацией, представленной в виде объектов – экземпляров понятий предметной области. Для описания данного процесса предложены три базовые модели. Модель представления текста задаёт общую схему обработки текста и обеспечивает отображение полученной информации на текст. Модель представления знаний включает описание предметной лексики, жанровые модели текста и модели фактов, которые позволяют смоделировать процессы извлечения информации в терминах семантических классов предметной лексики и онтологии предметной области. Используемая атрибутивная модель представления данных обеспечивает сохранение информационных потоков данных, возникающих в процессе извлечения информации, и позволяет применять онтологические методы для решения задач снятия неоднозначности интерпретации текста и разрешение кореференции. Таким образом, предложена оригинальная методика, позволяющая пользователям проектировать систему анализа текста и моделировать процессы извлечения информации на основе онтологии предметной области.

Ключевые слова: *извлечение информации, модель текста, словарь предметной лексики, модель факта, пополнение онтологии.*

Цитирование: *Сидорова, Е.А.* Подход к моделированию процесса извлечения информации из текста на основе онтологии / Е.А. Сидорова // Онтология проектирования. – 2018. – Т. 8, №1(27). - С.134-151. – DOI: 10.18287/2223-9537-2018-8-1-134-151.

Введение

Автоматическая обработка и анализ разнородной информации, представленной на естественном языке, является одним из самых востребованных на сегодняшний день направлений исследований. Отдельное положение в ряде решаемых задач в рамках данного направления занимает задача извлечения информации из предметно-специфического контента, представленного текстами, поскольку данная задача плохо разрешима статистическими методами и, как правило, требует привлечения знаний специалистов как о предметной области (ПрО), так и об особенностях языка.

Данная задача тесно связана с задачей пополнения онтологии [1]. Под пополнением онтологии понимается автоматический анализ различных источников и наполнение найденной информацией контента информационной системы, база данных которой опирается на онтологию ПрО. Извлекаемая информация в таких системах представляется экземплярами понятий и отношений заданной онтологии. Для решения данных задач, как правило, используют разнообразные знания в формализованном виде, такие как тезаурусы (WordNet, RusNet), толково-комбинаторные словари [2], аннотированные корпуса текстов (например, Национальный корпус русского языка - www.ruscorpora.ru) и т.п. Работа со знаниями в свою очередь требует создания технологий, которые автоматизируют процессы проектирования и разра-

ботки программных систем посредством предоставления пользователям (в том числе не программистам) средств моделирования, которые позволяют абстрагироваться от непосредственной разработки программных компонент и сконцентрироваться на вопросах обеспечения системы всеми необходимыми знаниями и моделирования непосредственно процессов извлечения информации в предметных терминах.

Вопросам моделирования процессов извлечения информация уделяется мало внимания, как правило, рассматриваются решения конкретных задач, что не позволяет выделить технологические и методологические аспекты решения подобных задач в целом. Целью данной работы является попытка восполнить этот пробел с помощью ряда модельных описаний процессов на различных уровнях представления.

Наиболее изученными являются онтологии и тезаурусы [3]. Онтология, как инструмент моделирования ПрО, является основой для формализации интересов пользователя, формирования словаря и представления конечного результата работы создаваемой системы, а также содержит необходимые знания для этапа семантического анализа текста [4]. Тезаурус, как инструмент описания предметной лексики, позволяет характеризовать термин и его связи с точки зрения особенностей употребления в данной ПрО [5, 6]. Для того, чтобы зафиксировать формально-лингвистические свойства, обусловленные языковой практикой описания объектов и ситуаций данной ПрО, необходимо использовать модели, которые бы с одной стороны описывали различные варианты представления в тексте одной и той же информации, с другой – моделировали бы процесс извлечения данной информации. К моделям данного класса можно отнести синтаксические или семантико-синтаксические модели управления [7, 8], лексико-синтаксические правила и шаблоны [9, 10], правила на основе онтологии [11] и т.п.

Указанные выше модели позволяют формировать в первую очередь базу знаний разрабатываемой системы, т.е. моделировать «заранее» заданные параметры системы или входные данные. Для моделирования целостного процесса извлечения информации их оказывается недостаточно, необходимо указать структуры данных для представления информации на всех промежуточных этапах обработки текста, а также обеспечить решения «специализированных» для данной проблемной области задач, связанных с неоднозначностью, присущей естественному языку.

В данной работе предложен подход к моделированию процессов извлечения информации из текста, который включает компоненты моделирования базы знаний системы, модели представления текста в процессе его обработки, а также способы описания информационных потоков данных, возникающих в процессе извлечения информации, которые позволяют применять онтологические методы для решения задач снятия неоднозначности и разрешение кореференции.

1 Модель представления знаний

Особенностью развиваемого подхода к извлечению информации из текста является применение знаний о ПрО, преимущественное использование лексико-семантической информации и жанровых особенностей документов.

Рассматриваемая лингвистическая модель знаний включает три компонента. Словарь ПрО задаёт лексическую модель подъязыка ПрО, жанровая модель текста формирует жанровую структура рассматриваемого текстового источника, сужая область поиска определённой информации, и модели фактов, связывающие семантико-синтаксические модели, описывающие структуру выражений, принятых в данной области для описания информации, с формальным представлением этой информации, определяемым онтологией ПрО.

1.1 Словарь предметной лексики

В рамках предлагаемого подхода в качестве лексической модели языка рассматриваются информационно-поисковые словари ПрО. Такого рода словари ориентированы на автоматическую обработку текста и содержат дополнительную информацию, позволяющую распознавать термины в текстах.

Формально предметный словарь определяется системой вида <*V, M, T, S*>, где

$V = W \cup P \cup L$ – множество предметных терминов, включающих:

- W – множество лексем (каждой лексеме сопоставлена информацию обо всей совокупности её форм;
- P – множество словокомплексов или многословных терминов, характеризующихся высокой частотностью в анализируемом подъязыке (словокомплекс описывается парой <L-грамма, тип структуры>, где L-грамма задаёт последовательность лексем, а тип структуры определяет вершину и правила согласования элементов L-граммы);
- L – множество лексических конструкций, каждая из которых описывается с помощью шаблонов, используемых для распознавания регулярных текстовых фрагментов (лексические конструкции предназначены для распознавания таких структур как сокращения, аббревиатуры, численные или буквенно-численные обозначения объектов ПрО или значения их атрибутов).

M – морфологическая модель языка, включающая описание морфологических классов и атрибутов (атрибуты в рамках каждого класса делятся на словообразующие, присущие всем формам лексемы данного класса, и словоизменительные, различающие формы одной лексемы).

T – множество тематических признаков, организованных в иерархию (каждому термину может быть сопоставлен набор признаков с указанием веса связи, где вес – это значение из интервала [0,1], отражает степень принадлежности термина признаку).

S – лексико-семантическая модель ПрО.

Рассмотрим подробнее последний компонент, который является особенно важным при моделировании процесса извлечения информации.

Модель предметной лексики должна включать описание структуры семантики терминов и позволять, в конечном итоге, сопоставлять текстовым единицам их смысловые эквиваленты. Предложенная модель включает грамматическую, тезаурусную и семантическую информацию о термине, а также необходимые данные для описания валентностной структуры предикатных слов.

Для кодирования семантической информации о слове предусмотрены следующие возможности (см. рисунок 1).

Семантический класс. Термин может быть отнесён к определённому семантическому классу. Иерархия классов позволяет отнести термин к определённому уровню иерархии, более общему или конкретному с наследованием свойств общего класса.

Семантический атрибут. Для представления лексического значения термина используются семантические атрибуты. Совокупность значений атрибутов, приписанных слову, в определённой мере моделирует компонентную семантическую структуру слова. Основные компоненты семантической структуры термина могут рассматриваться как тезаурусные дескрипторы.

Группировка семантических классов и атрибутов для описания многозначного слова. Если термин имеет более одного формально различимого контекстом значения, формируется соответствующее число семантических статей слова, объединяющих в себе с помощью механизма группировки семантический класс и совокупность семантических атрибутов с их значениями.

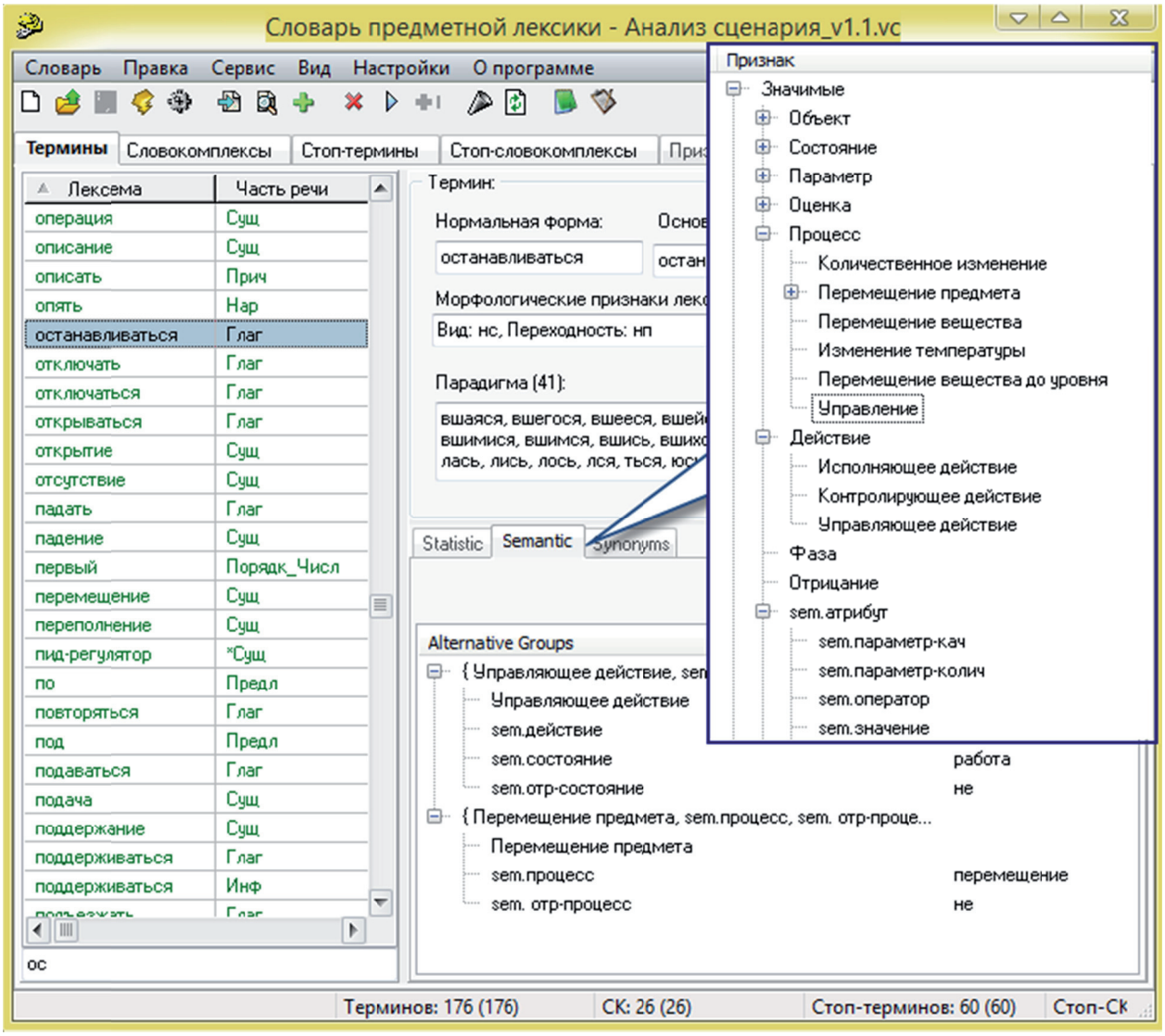


Рисунок 1 – Пример описания семантики термина

Семантика найденного в тексте термина обеспечивает формирование лексического объекта. Структуру лексического объекта можно представить следующим образом:

```
LexObject
Name: string;                  // название термина
Descriptor: string;            // имя дескриптора
Semantic: set of LexClass;     // множество лексико-семантических классов
Neg: bool;                     // наличие отрицания
Attributes: set of Semantic_attribute;        // множество атрибутов
Value: string;                 // значение или форма, в которой встретился термин в тексте
Position: int;                 // позиция термина в тексте
Grammatics: set of gram_parameters;           // грамматические характеристики термина
```

Параметр *Name* задаёт нормализованную форму термина, полученную либо в соответствии с правилами морфологической нормализации (в случае однословного термина), либо в соответствии с правилами согласования слов в словосочетании (для многословных терминов), либо как имя лексического шаблона (для символьно-числовых конструкций). Параметр

Descriptor определяет имя понятия, к которому относится термин. В частности, дескриптор служит для формирования синсетов или групп синонимов, когда все термины с одинаковым дескриптором являются синонимами или квазисинонимами в данной ПрО. Поле *Semantic* определяет лексико-семантический класс(ы) термина. Наличие нескольких значений в поле *Semantic* лексического объекта соответствует составному значению термина (например, сложносоставное слово *пятиметровый* имеет лексическое значение *Числа* и *Единицы измерения*) в отличие от неоднозначности, когда формируется несколько лексических объектов *LexObject* (например, по термину *останавливаться* будут созданы объекты с семантикой *Действия* и *Процесса*). Отметим, что морфологическая омонимия не порождает отдельных лексических объектов, а разрешается, если необходимо, в рамках одного объекта. Поле *Neg* задаёт наличие отрицания при значении термина, которое в тексте может выражаться, например, приставкой *не*. Поле *Attributes* задаёт множество семантических атрибутов для представления структуры семантики термина. Поля *Grammatics*, *Position* и *Value* определяются словарными и текстовыми характеристиками термина.

1.2 Жанровая модель текста

Предлагаемые решения по анализу и извлечению информации из текстов опираются на понятие жанра как совокупности содержательных и формально-лингвистических (логико-композиционных и лексико-грамматических) аспектов.

Жанровая модель текста представляется системой вида GMT = $<g, F_S, LC, V_G, M_G, P_G>$, где:

g – жанровый тип документа;

F_S – множество формальных жанровых сегментов (каждый сегмент, найденный в тексте в соответствии с жанровой моделью, определяется типом $t_f \in F_S$ и начальной и конечной текстовой позицией);

LC – логико-композиционная структура текста, определяемая множеством взаимосвязей между текстовыми фрагментами;

V_G – словарь жанровой лексики;

M_G – множество жанровых маркеров, которые задаются с помощью терминов из V_G;

P_G: $M_G \rightarrow F_S$ – множество жанровых шаблонов, которые связывают маркеры и структурные блоки текста.

Таким образом, жанровые особенности текста передаются его разбиением на содержательные блоки, которые:

- включают определённую жанровую лексику,
- имеют определённую структурную организацию,
- реализуются в рамках определённых формальных сегментов.

Логико-композиционная структура текста выявляется с помощью лексикона жанровых маркеров и шаблонов, выделяющих содержательные блоки. Маркеры извлекаются из текстов, и прежде всего – из заголовков подразделов, вводных предложений и списков, включенных в состав документа. Простые маркеры сопоставляются терминам предметного словаря или группе терминов (синонимы). Более сложные формируются на основе простых: поддерживаются альтернативы, совместная встречаемость, а также вложенное использование маркеров.

На рисунке 2 в качестве примера приведён фрагмент жанровой модели текста стандартного протокола клинических испытаний. Данные тексты включают жанровые сегменты, маркированные жанровыми тегами, на основе которых могут извлекаться фрагменты текста для поиска той или иной информации.

Выявленные на основе жанровых моделей значимые сегменты могут в дальнейшем использоваться в качестве условия поиска определённой информации.

TextGenre **Block** genre_segment **Block** genre_segment … **Block** genre_segment **Block** genre_segment …	*<study_design_info>* *<allocation>Randomized</allocation>* *<intervention_model>Parallel Assignment</intervention_model>* *<primary_purpose>Prevention</primary_purpose>* *<masking>Double Blind (Participant, Care Provider, Investigator, Outcomes Assessor)</masking>* *</study_design_info>*

Рисунок 2 – Фрагмент жанровой модели текста, описывающего дизайн клинического исследования

1.3 Модель факта

Модель факта является основой для моделирования процесса извлечения информации. Каждая модель задаёт схему извлечения единицы информации или единичного факта, представленного связанным языковым выражением. При извлечении факта из текста необходимо учитывать множество языковых способов репрезентации данного факта носителями подъязыка и обеспечивать их трансформацию в формальную структуру. Совокупность моделей должна описывать процесс целиком, начиная с инициализации объектов, заполнения атрибутов и установления онтологических отношений между объектами.

Факт, представляя собой зафиксированное в высказывании (языковом выражении) эмпирическое знание об объектах, их свойствах и ситуациях, может быть формализован в виде когнитивной схемы, соотносящей его с понятиями и отношениями онтологии. Таким образом, модель фактов формирует знания о согласовании имеющихся лингвистических знаний с предметными знаниями и фиксирует формально-лингвистические свойства, обусловленные языковой практикой описания объектов и ситуаций данной ПрО.

Предлагаемый способ моделирования обеспечивает преимущественное использование лексико-семантической информации, что не исключает применения частичного синтаксического анализа и синтаксических ограничений, накладываемых на семантический каркас моделей фактов. Известные системы отличаются полнотой и ролью синтаксического анализа в процессе извлечения фактической информации из текстов. Так, технология [12] предполагает построение полного семантико-синтаксического дерева предложения, к которому применяются шаблоны (своего рода фильтры), описывающие искомые факты. В предлагаемом подходе, как и в [13], синтаксический анализатор применяется локально (при обнаружении ключевых единиц и их конфигураций), в частности, предусмотрено определение актантных позиций предикатных слов [14].

Основой лингвистического описания факта является семантико-синтаксическая модель, которая ограничивает синтаксическую сочетаемость и согласованность грамматических и семантических признаков терминов (вершин синтаксических групп) в соответствии с правилами согласования и управления. Такие модели описываются в виде актантной структуры, связанной с одной или несколькими обобщёнными лексемами [9, 15]. Под обобщённой лексемой понимается либо термин словаря (или его форма), либо группа лексем, описанных в терминах грамматических и семантических категорий без указания нормальной формы. Актантная структура описывает набор актантов, характеризующих соответствующую валентность, в терминах семантических и грамматических характеристик, которые являются ограничениями для зависимых слов.

Семантико-синтаксическая модель характеризуется парой вида $<lg, M_A>$, где:

$lg = <L_{lg}, S_{lg}, M_{lg}>$ – обобщённая лексема, характеризующая группу терминов $L_{lg} \subseteq V$, обладающую набором семантических признаков $S_{lg} \subseteq S$ и заданными значениями морфологических атрибутов $M_{lg} \subseteq Mv$;

$M_A = <A_1, ..., A_n>$ – последовательность актантов, описывающих модель, где каждый актант $A_i \subseteq S \times Mv \times V_{pre}$ задаёт множество допустимых сочетаний семантических S и грамматических Mv характеристик терминов, а также условия на наличие предлога V_{pre}.

Предложенная структура семантико-синтаксических моделей предоставляет широкие возможности моделирования языковых связей в тексте. Так, модель может не содержать синтаксических ограничений и представлять собой онтологические отношения или описываться без семантических характеристик и соответствовать чисто синтаксическим моделям управления. Обобщение лексем в моделях позволяет компактно определить многие языковые конструкции, варианты взаимосвязи слов в выражениях и словарные группы.

Модель фактов задаётся структурой, аналогичной актантной структуре, которая описывается либо в терминах классов онтологии, либо в терминах семантических признаков словаря и связывается с фрагментом онтологии. Дополнительно накладываются ограничения на онтологические признаки элементов структуры и их взаиморасположение в тексте.

Формально модель фактов – это система вида $F = <M_{AF}, M_{SS}, Cs, C_P, O_{Res}>$, где:

M_{AF} – последовательность аргументов факта;

M_{SS} – множество семантико-синтаксических моделей, которым должны удовлетворять аргументы на уровне лексического состава;

Cs – множество семантических ограничений;

C_P – множество структурных ограничений;

O_{Res} – фрагмент онтологии, в соответствии с которым формируется результат применения модели фактов - новые объекты и/или изменения атрибутов уже существующих объектов, которые являются экземплярами понятий онтологии (относящихся к заданному фрагменту).

Выделяются два основных типа моделей фактов: модели, служащие для начальной инициализации объектов, и модели для выявления связей. Модели первой группы необходимы для начального формирования онтологических сущностей на основании словарных признаков. Модели второй группы моделируют процессы «обнаружения» фрагментов онтологии.

Рассмотрим эти два типа на примерах.

Система семантических признаков словаря формируется на основе онтологических сущностей, что позволяет инициализировать начальное формирование объектов непосредственно на основании словарных признаков. Рассмотрим пример описания модели для инициализации объектов класса *Препарат*, которые могут быть представлены в тексте аппозитивной именной группой. В этой группе опорным словом является родовое слово или словокомплекс (тип), а имя примыкает к нему в постпозиции.

Scheme Препарат3 : segment *Клауза*
arg1: Term::Препарат(SemClass: тип)
arg2: Term:: Препарат(SemClass: имя)
Condition PrePos(arg1,arg2), Contact(arg1,arg2)
⇨ Object :: Препарат(Тип: arg1.Class & arg2.Class, Наименование: arg2.Norm)

В данной модели термины должны иметь семантических класс *Препарат*, с учётом иерархии наследования признаков в словаре, а также первый термин должен обладать семантическим признаком *тип*, а второй – *имя*. При применении данной модели, например, к фразе:

вакцина <Препарат, SemClass: тип> *"АСАМ2000"* <Препарат, SemClass: имя> [1]
создаётся объект – экземпляр понятия онтологии *Препарат*, тип препарата (например, фармакологическая группа) может уточниться в соответствии с семантическим признаком первого или второго термина, атрибут *Наименование* у объекта заполняется наименованием второго термина *Norm*.

В качестве другого примера инициализирующей модели рассмотрим случай, когда на основе лексического шаблона (LexTerm) выделяется фрагмент текста в кавычках и формируется гипотеза о том, что это имя объекта, но уточнение его класса возможно только при наличии термина-классификатора:

Scheme Новый_объект : segment *Клауза*
arg1: Term:: (SemClass: тип)
arg2: LexTerm::Именованный объект()
Condition PrePos(arg1,arg2), Contact(arg1,arg2)
⇨ Object (Тип: arg1.Class, Наименование: arg2.Name)

Формирование объекта с помощью данной модели осуществляется аналогично предыдущей.

При поиске и выявлении характеристик объектов и их связей, как правило, требуется проверить сочетаемость семантических и/или грамматических признаков объектов.

Рассмотрим примеры модели, используемой для извлечения атрибутов объектов.

Scheme ТипКогорты: genre_segment <*arm_group*>
arg1: Object:: Группа(), genre_segment <*arm_group_label*>
arg2: Term::Параметр(), genre_segment <*arm_group_type*>
⇨ arg1: Группа (тип: arg2.Name)

Данная модель позволяет уточнить тип группы на основе информации о жанровой структуре документа, которая в соответствии с принятым стандартом содержится строго в определённых жанровых фрагментах.

Рассмотрим пример модели для построения отношения в соответствии с рассматриваемой ситуацией.

Scheme УсловияПримененияПрепарата: genre_segment <*group*>
arg1: Object::Группа()
arg2: Object::Препарат()
arg3: Term::Параметр(SemClass: кратность)
arg4: LexTerm::Доза()
arg5*: Term::Параметр(SemClass: время)
Condition genre_segment (arg2, arg3, arg4, arg5) <*description*>,
Contact(arg2, arg3), Contact_weak(arg2, arg4), Contact_weak(arg2, arg5)
⇨ имеетНазначение (группа: arg1, препарат: arg2, тип: «препарат»,
размер дозы: arg4.value, кратность: arg3.value, время: arg5.value)

Условия назначения препарата для конкретной группы людей описываются такими характеристиками, как наименование препарата, его дозировка, кратность применения и время приёма. Данная информация в соответствии с принятым стандартом содержится строго в определённых жанровых фрагментах, однако в рамках фрагмента *description* описание пара-

[1] В примерах в скобках указываются признаки терминов, заданные в словаре.

метров разворачивается в виде текста, что требует применения более сложного лингвистического анализа. В результате применения модели создается описание ситуации терапевтического вмешательства.

Приведённый набор моделей фактов демонстрирует подход к извлечению информации о проводимом клиническом исследовании на основе структуры протокола.

2 Модель представления текста

Важным компонентом моделирования процесса извлечения информации является модель представления текста, которая последовательно изменяется, обогащаясь на каждом этапе анализа новыми знаниями. Для описания происходящих изменений предложена модель, близкая по смыслу к схемам, используемым при создании размеченных корпусов текстов [16]. Отличия заключаются в:

- единообразной поддержке всех этапов обработки текста, включая семантический;
- использовании «внешнего» аннотирования (а не систему тэгов), синхронизированного с текстом [17];
- ориентации на объектно-ориентированное представление данных.

Модель представлена набором *покрытий текста*, когда промежуточные результаты обработки представляются однотипными объектами с заданной проекцией на текст (текстовыми интервалами), что позволяет наглядно интерпретировать полученные результаты и выделять контекстно связанные с каждым элементом знания.

Модель текста определяется пятеркой $<C_A, C_L, C_G, C_{Th}, C_{IO}>$, где:

$C_A = \{atom_1, ..., atom_n\}$ – графематическое покрытие, содержащее множество атомов $atom_i = <t, id, pos>$, где атом – это объект, сопоставляемый неразрывному фрагменту текста, состоящему из символов одного типа t; атомы упорядочены по встречаемости в тексте и определяются порядковым номером id и текстовыми позициями $pos = <pos\_begin, pos\_end>$;

$C_L = \{lex_1, ..., lex_n\}$ – терминологическое покрытие, содержащее множество лексических объектов вида $lex_i = <v, m_v, s_v, pos>$, где $v \in V$ - термин словаря; m_v – множество грамматических характеристик термина v; s_v – множество семантических признаков v; pos – текстовая позиция;

$C_G = \{s_1, ..., s_n\}$ – сегментное (или жанровое) покрытие, отражающее логико-композиционную структуру текста и включающее множество сегментов вида $s_i = <t_f, pos, R_G>$, где t_f – тип или формальный сегмент жанровой модели текста, pos – текстовая позиция и R_G – связи с другими сегментами, определяющими их взаиморасположение в тексте;

$C_{Th} = \{st_1, ..., st_n\}$ – тематическое покрытие, которое определяется множеством тематических фрагментов вида $st_i = <th, pos, R_{Th}>$, где th – тематика из заранее определённого тематического классификатора, pos – текстовая позиция и $R_{Th} \subseteq C_G$ – подмножество структурных сегментов, покрываемых данным фрагментом;

$C_{IO} = \{IO_1, ..., IO_n\}$ – информационное покрытие, содержащее множество найденных в тексте информационных объектов *вида* $IO_i = <I, Pos, R_I>$, где I – онтологический объект или экземпляр понятия, определённого онтологией ПрО, Pos – текстовая позиция (в общем случае разрывная, т.е. определяемая множеством неразрывных позиций pos), R_I – множество информационных зависимостей объекта, полученных в процессе обработки текста при использовании информации из одного объекта для генерации или обновления другого.

В зависимости от решаемой задачи могут быть выделены и другие типы покрытий. Приведённая модель ориентирована в первую очередь на задачи семантического анализа и извлечения информации. На рисунке 3 представлена общая схема преобразования входящих текстовых данных и промежуточные результаты в виде покрытий.

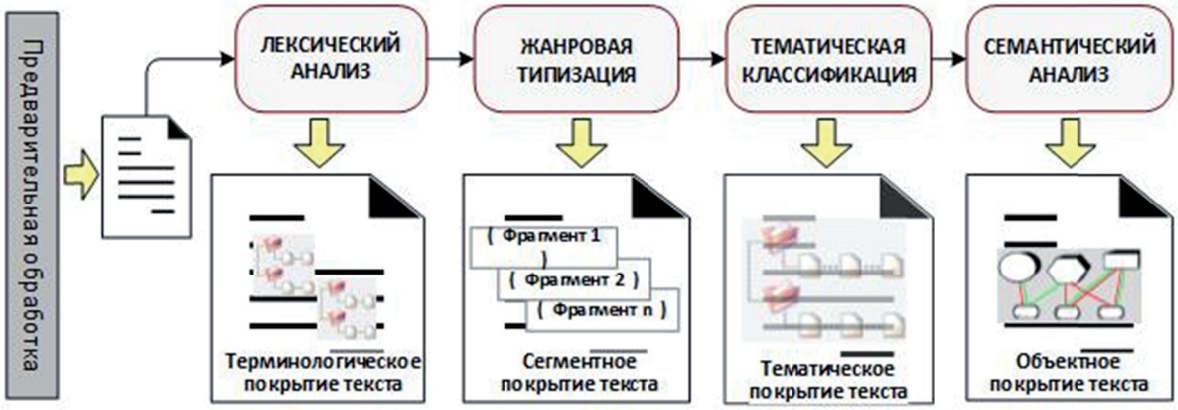


Рисунок 3 – Этапы формирования модели представления текста

Графематическое покрытие текста является результатом его графематического анализа, в процессе которого входящий линейный текст разбивается на элементарные атомы. Основная задача данного этапа – сгруппировать символы одного типа в последовательности и дать им необходимую интерпретацию: слово определённого алфавита, число, знак препинания и т.п. Для решателей, работающих с разметкой (например, html-тексты), можно дополнительно задать типизацию тэгов или помет. Важным свойством данного представления является то, что элементы покрытия задают все возможные границы элементов для всех последующих представлений, т.е. при дальнейшей обработке ни один атом не может быть «разделён».

Терминологическое покрытие состоит из словарных терминов, найденных в данном тексте, с учетом возможной омонимии и пересечений многословных терминов. Терминологическое покрытие текста – это лексическая модель текста, которая строится на основе лексической модели подъязыка ПрО, и включает найденные в тексте термины с привязкой к позиции в тексте. После того, как термин найден в тексте, формируется лексический объект, который снабжается набором семантических атрибутов, заданных в предметном словаре для найденного термина (см. раздел 1.1).

Сегментное покрытие отражает структурное деление текста на логические (абзац, предложение, заголовок и т.п.) и жанровые фрагменты. Жанровое покрытие является результатом сегментации текста и одним из способов отражения его формальной структуры. Сегментация рассматривается на макроуровне, т.е. на уровне всего текста (в отличие от локального анализа предложения и выделения совокупности взаимосвязанных фрагментов (клауз), рассматриваемых в рамках синтаксического анализа предложения) и опирается как на формально-текстовые, так и на жанровые особенности документа (см. раздел 1.2), которые передаются разбиением текста на концептуальные части. При анализе текста разбиение на жанровые фрагменты помогает сузить область поиска информации определённого вида и, тем самым, повысить качество анализа. С помощью жанровой структуры текста могут решаться задачи определения жанровой релевантности документов, полученных из неизвестных источников, например, при поиске в Интернет [18].

Тематическое покрытие определяет текстовые границы тематически-связанных областей текста для каждой рассматриваемой тематики. Формирование таких областей осуществляется на основе словаря, в котором задано соответствие между терминами и тематическими признаками. Тематическое покрытие строится над терминологическим и жанровым покры-

тием. Здесь элемент тематического покрытия определяется как фрагмент текста, включающего кластер терминов, относящихся к одной теме, в границах формального сегмента (или последовательности сегментов) жанрового покрытия. Аналогично сегментам, тематические сегменты могут сужать область поиска информации определённого вида.

Информационное (объектное) покрытие описывает найденную информацию в виде семантической сети объектов ПрО. Информационное покрытие текста является самым информационно-насыщенным и представляет результаты семантической обработки документа. Чтобы построить информационное покрытие необходимо представить содержание документа в понятной компьютеру форме. Условием этого является наличие формата данных, задающего строгую структуру представления и хранения полученной информации. Данная структура должна быть «заранее» осмыслена, т.е. иметь заранее заданную семантическую интерпретацию. В современных подходах для этого используют онтологии ПрО [19].

В упрощённом виде онтологию можно представить как $O = <C, T, A>$, где:

C – множество классов, описывающих понятия ПрО;

$T = \cup T_i$ – множество типов данных и V_i – множество значений типа T_i;

A – множество атрибутов, $F_A: C \rightarrow 2^{A \times (T \cup C)}$ – функция, которая определяет имена и типы атрибутов классов C; выделяются атрибуты простых типов (owl:DataProperty) и связи (owl:ObjectProperty), также определяется подмножество ключевых атрибутов $A_K \subseteq A$, которые служат для однозначной идентификации объектов.

Контент информационной системы, построенной на основе онтологии O, представляется множеством экземпляров классов онтологии и описывается как $I = \{I_1,..,I_n\}$, где I_i класса C_i представляется набором атрибутов со значениями $Av_i = \{(a_1,v_1), ..., (a_k,v_k)\}$.

Объектное покрытие описывает информационный контент текста в терминах информационных объектов, которые формируются на основе онтологических понятий и должны быть сопоставлены определённому экземпляру онтологии, который либо уже присутствует в контенте информационной системы, либо будет добавлен в результате анализа текста. Таким образом, информационные объекты являются своего рода «вхождениями» или упоминаниями онтологических объектов, обнаруженными в тексте.

Информационные объекты формируются на основе моделей фактов (см. раздел 1.3.), при этом порождаются информационные зависимости между объектами, выступающими в качестве аргументов модели, и её результатом. Для точного описания данных зависимостей используется атрибутивная модель извлечения информации.

Преимущество предложенной модели текста заключается: во-первых, в наглядном представлении результатов работы анализатора; во-вторых, предложенное описание может являться основой для формального описания алгоритмов и доказательства их свойств, а также служить в качестве абстракции верхнего уровня для программной реализации; в-третьих, использование данного представления в рамках информационных систем обеспечит достоверность результата, подтверждаемого непосредственно текстовым источником, что позволит проводить широкий спектр корпусных исследований.

3 Атрибутивная модель извлечения информации

При описании последовательных процессов извлечения информации с помощью моделей фактов интенсивно используются атрибуты объектов и связи между ними. Поэтому для детального представления происходящих процессов и отражения возникающих информационных зависимостей не только на уровне объектов, но и между их атрибутами, решено воспользоваться подходом к моделированию процессов на основе атрибутной конвейерной модели (АКМ).

Атрибутная модель – это логическая модель данных, предназначенная для отображения связей между свойствами или атрибутами объектов, участвующих в одном процессе. Для корректного моделирования процессов извлечения информации модели должны обладать свойством конвейерности, т.е. позволять описывать процессы, внутри которых не возникают циклы. Такого рода модели широко используются в различных ПрО, например, при описании технологических процессов с помощью диаграмм сборки Сервис-Компонентной Архитектуры SCA [20]. Использование АКМ позволяет выстраивать общую схему процесса с зависимостями по атрибутам.

В общем виде модель АКМ можно описать системой вида $<O_M, A, L>$, где:
O_M – множество объектов моделирования;
A – множество атрибутов объектов (подразумеваются атрибуты простых типов данных);
L – множество направленных связей между атрибутами объектов.

При моделировании процесса извлечения информации связь между атрибутами интерпретируется как передача значения от одного атрибута к другому. Для обеспечения целостности модели на связи накладываются следующие ограничения:

- атрибут может иметь только одну входящую связь и множество исходящих связей;
- связь можно установить только между атрибутами одного типа.

Для поддержки моделирования процессов извлечения информации на основе онтологии и моделей фактов было необходимо расширить АКМ таким образом, чтобы она поддерживала следующие возможности:

- в качестве атрибута объекта может выступать другой объект, и, соответственно, модель должна предоставлять возможность устанавливать связь между объектом и атрибутом другого объекта (соответствующего типа);
- возможность устанавливать значение атрибута напрямую без использования связей, т.е. без передачи значения от другого объекта;
- в случае порождения нового объекта (в соответствии с моделями фактов) возможность устанавливать связи между аргументами факта и порождаемым объектом.

Для реализации первой возможности была расширена типизация свойств АКМ – добавлена возможность использовать класс онтологии в качестве типа данных атрибута (аналогично онтологическому объекту), также в модель введены соответствующие связи между объектом и атрибутом, которые обладают теми же возможностями и ограничениями, что и исходящие связи. Для поддержки второй возможности были добавлены *независимые атрибуты* – атрибуты, значения которых можно изменять напрямую, не используя соединения (при этом возникают неявные зависимости между аргументами факта и независимым атрибутом, которые пока не моделируются). И, наконец, в случае если других связей не установлено, между объектом-аргументом и результирующим объектом устанавливается *объектная связь*, которая реализуется на уровне имён классов, т.е. имя класса объекта в данной модели рассматривается как его атрибут.

Предложенная модель имеет два практических применения. В первом случае АКМ является основой (метамоделью) для моделирования процессов с помощью моделей фактов: как для описания единичных моделей (см. рисунок 4), так и для их совокупности. В другом случае в процессе обработки текста она используется для сохранения информационных зависимостей между объектами информационного покрытия текста (данные информационные зависимости согласуются с зависимостями, заданными в моделях фактов) и, тем самым, представляет «историю» создания объектов. Эта информация позволяет, в частности, оценить степень связности объекта с контекстом и осуществить корректное удаление объекта из системы.

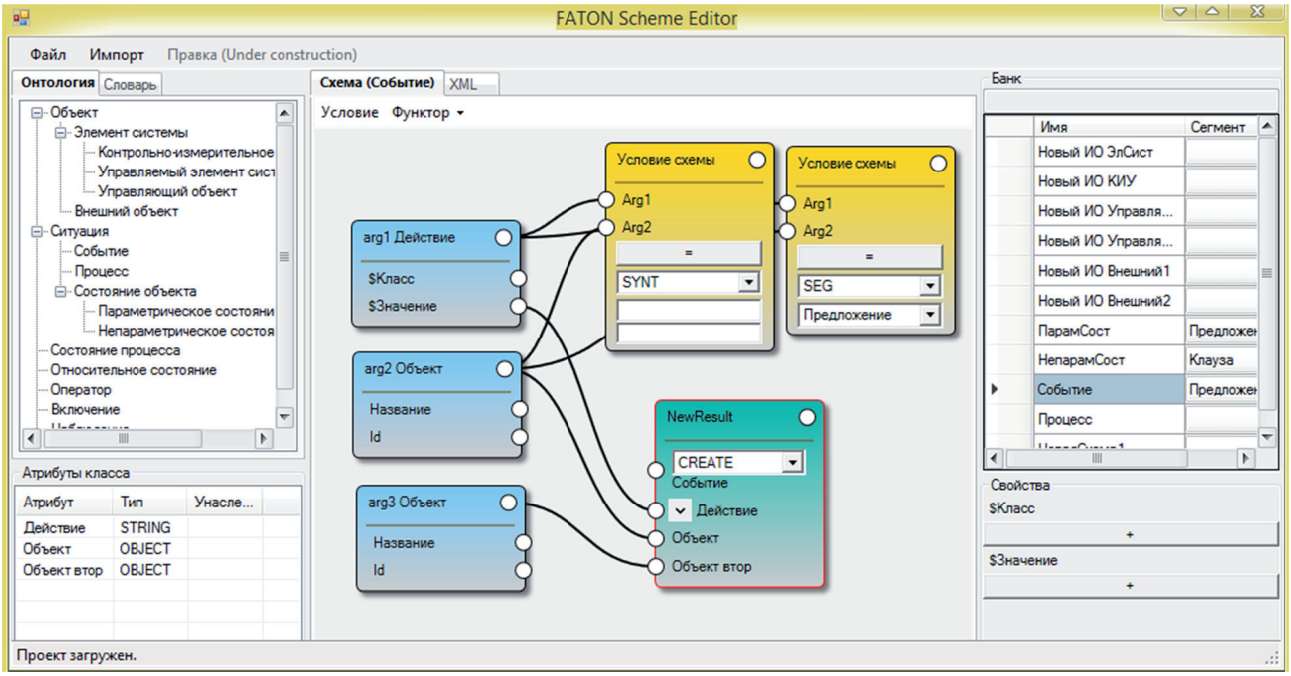


Рисунок 4 – Пример использования АКМ для моделирования процесса извлечения информации о единичном факте

4 Извлечение информации

Задача извлечения информации в соответствии с предложенными моделями представления текста сформулирована в общем виде следующим образом.

Для заданной четверки <*O, I, LM, T*>, где: *O* – онтология ПрО, *I* – информационный контент системы, *LM* – модель подъязыка ПрО, *T* – текстовый источник, построить информационное покрытие текста C_{IO}, обладающее свойством однозначности и сопоставимости с *I*. Под однозначностью понимается наличие не более одной интерпретации любого фрагмента заданного текста *T*. Это означает, что при наличии нескольких вариантов разбора текста, например, в случае омонимии, должен быть выбран один вариант. Свойство сопоставимости с контентом системы означает, что каждому информационному объекту, входящему в результирующее информационное покрытие C_{IO}, должен быть однозначно сопоставлен экземпляр онтологии O, что обеспечивается наличием заданных ключевых атрибутов. Сопоставляемый экземпляр, по сути, является онтологическим референтом найденного в тексте объекта.

На рисунке 5 представлена общая схема реализации семантического этапа анализа и извлечения информации из текста. На вход системы основного анализа поступают результаты предварительного этапа обработки текста в виде терминологического, сегментного и тематического покрытия текста. База знаний системы включает знания об онтологии ПрО и моделях фактов, заданных для данной онтологии. Результатом работы системы является объектное покрытие текста, на основе которого формируется результирующее множество онтологических объектов, описывающих контент документа в терминах онтологии ПрО.

В процессе основного анализа можно выделить следующие три основных задачи:

1) Непосредственно извлечение фактов на основе моделей фактов.
2) Разрешение кореференции и поиск референтов объектов в онтологическом контенте системы.
3) Разрешение неоднозначности.

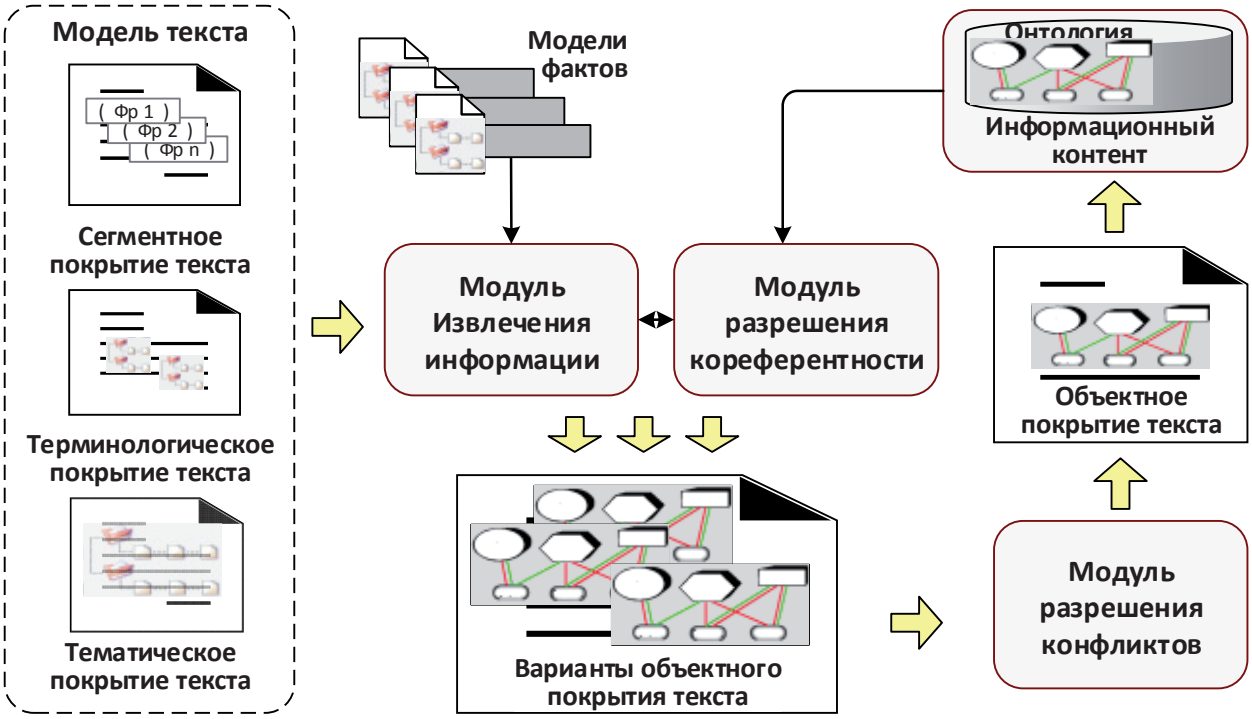


Рисунок 5 – Схема извлечения информации из текста

Решение первой задачи моделируется пользователем при разработке моделей фактов. Технологическая поддержка данного процесса осуществляется системой, которая по моделям фактов формирует правила и применяет их к входным данным (одна из возможных реализаций данной системы представлена в [21]). Качество извлечения информации в большой степени зависит от качества модели процесса, созданной пользователем. В результате формируется АКМ процесса извлечения информации для данного текста.

От решения второй задачи во многом зависит полнота анализа. В работе [22] предложен метод разрешения кореференции на основе онтологии, в рамках которого осуществляется семантическое сравнение найденных в текстах объектов с учётом их онтологических свойств. На данном этапе ищутся все упоминания каждого онтологического объекта, формируются кореферентные группы информационных объектов и разрешаются возможные конфликты. На качество решения данной задачи влияют степень проработанности заданной пользователем онтологии и наполненность информационного контента системы.

Решение третьей задачи обеспечивает формирование конечного варианта объектного покрытия текста для размещения в информационном контенте системы. Неоднозначности при анализе текста возникают вследствие особенностей естественного языка. Языковая неоднозначность — это способность слова или выражения иметь различные интерпретации, в результате которой порождаются различные варианты разбора текста, конфликтующие между собой. В разрабатываемом подходе конфликты рассматриваются на уровне проекции результатов анализа текста на ПрО, т.е. в контексте заданной онтологии. Полученная при извлечении фактов АКМ характеризует информационные зависимости, возникшие в процессе формирования данной модели. Предлагаемый метод основан на идее вычисления степени информационной связанности информационных объектов, извлечённых из заданного текста [23]. Модуль разрешения конфликтов должен разрешить все неоднозначности таким обра-

зом, чтобы система была свободной от конфликтов и при этом сохранила максимально возможное количество объектов и связей.

Результат работы системы анализа помещается в информационное хранилище, предварительно обеспечив однозначную идентификацию найденных объектов методом, предложенным в работе [24].

Заключение

Представленный в работе подход к моделированию процессов извлечения информации существенным образом опирается на знания о ПрО, явно формализованные в виде онтологии, что позволяет применять методы локального семантического и синтаксического анализа, не требуя наличия полного корректного синтаксического разбора и грамматически правильно построенного текста. Сужение области значения предметных терминов значительно уменьшает неоднозначность текста. Использование информационного контента онтологии при идентификации и сравнении объектов, найденных в тексте, позволяет использовать неявные знания, т.е. информацию, не содержащуюся в тексте.

Рассмотренные модели и их технологическое обеспечение предоставляют конечным пользователям – экспертам в ПрО, лингвистам и инженерам знаний – инструменты для моделирования процессов извлечения информации и их отладки, а разработчикам информационных систем – инструменты для проектирования систем автоматической обработки текста и концептуальные схемы представления данных.

Благодарности

Работа выполнена при финансовой поддержке Президиума СО РАН (Блок 36.1. Комплексной программы ФНИ СО РАН II.1) и РФФИ (грант № 17-07-01600).

Список источников

[1] ***Petasis, G.*** Ontology Population and Enrichment: State of the Art / G. Petasis, V. Karkaletsis, G. Paliouras, A. Krithara, E. Zavitsanos // In Knowledge-driven multimedia information extraction and ontology evolution. - LNAI 6050. - Springer-Verlag Berlin, 2011. - P.134–166.

[2] ***Мельчук, И.А.*** Опыт теории лингвистических моделей: «Смысл-Текст». Семантика, синтаксис / И.А. Мельчук. – М.: Школа «Языки русской культуры», 1999. – 992 с.

[3] ***Нариньяни, А.С.*** ТЕОН-2: от Тезауруса к Онтологии и обратно / А.С. Нариньяни // Труды международного семинара Диалог'2002 по компьютерной лингвистике и ее приложениям. –М.: Наука, 2002. – Т.1. – С.307–313.

[4] ***Загорулько, Ю.А.*** Семантическая технология разработки интеллектуальных систем, ориентированная на экспертов предметной области / Ю.А. Загорулько // Онтология проектирования. - 2015. - Т.5. - №1 (15). – С.30-46.

[5] ***Добров, Б.В.*** Онтологии и тезаурусы: модели, инструменты, приложения: учебное пособие / Б. В. Добров, В. В. Иванов, Н.В. Лукашевич, В.Д. Соловьев. - М.: Интернет-университет информационных технологий; БИНОМ. Лаборатория знаний, 2009. - 173 с.

[6] ***Рубашкин, В.Ш.*** Семантический компонент в системах понимания текста / В.Ш. Рубашкин // Труды Десятой национальной конференции по искусственному интеллекту с международным участием КИИ-2006. – М.: Физматлит, 2006. – Т.2. – С.455-463.

[7] ***Розенталь, Д.Э***. Управление в русском языке / Д.Э. Розенталь. – М.: Книга, 1986. – 173 с.

[8] ***Журавлев, А.О.*** Создание базы данных моделей управления слов русского языка / А.О. Журавлев // Искусственный интеллект. 2013. № 1. – С.91-97.

[9] ***Большакова Е.И.*** Язык лексико-синтаксических шаблонов LSPL: опыт использования и пути развития / Е.И. Большакова // Программные системы и инструменты: Тематический сборник. - №15. – М.: МАКС Пресс, 2014. – http://www.lspl.ru/articles/Paper_19_LSPL.pdf.

[10] ***Ковалев, А.И.*** Инструмент разработки предметных словарей на основе лексических шаблонов DigLex / А.И. Ковалев, Е.А. Сидорова // Материалы Всероссийской конференции с международным участием «Знания – Онтологии – Теории» (ЗОНТ–2015), 6 - 8 октября 2015 г. – Новосибирск: Институт математики им. С.Л. Соболева СО РАН, 2015. – Т1. – С.123-130.

[11] ***Хорошевский, В.Ф***. OntosMiner: Семейство систем извлечения информации из мультиязычных коллекций документов / В.Ф. Хорошевский // Девятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2004: Труды конференции. В 3-х т. М.: Физматлит, 2004, т. 2. - С.573-581.

[12] ***Ермаков, А.Е.*** Автоматическое извлечение фактов из текстов досье: опыт установления анафорических связей / А.Е. Ермаков // Труды международной конференции Диалог’ 2007 «Компьютерная лингвистика и интеллектуальные технологии». - М.: Наука, 2007. – С.131–135.

[13] ***Гершензон, Л.М.*** Система извлечения и поиска структурированной информации из больших текстовых массивов СМИ. Архитектурные и лингвистические особенности / Л.М. Гершензон, И.М. Ножов, Д.В. Панкратов // Труды международной конференции Диалог’2005 «Компьютерная лингвистика и интеллектуальные технологии». М.: Наука, 2005. - С.97-101.

[14] ***Азарова, И.В.*** Использование маркеров актантных позиций при анализе деловых текстов для расширения логической схемы ПрО / И.В. Азарова, А.С. Гребеньков, Т.М. Ландо // Труды международной конференции Диалог’ 2008 «Компьютерная лингвистика и интеллектуальные технологии». - М.: РГГУ, 2008. - Вып. 7 (14). - С.11-16.

[15] ***Яковчук, Е.И.*** Обобщенные семантико-синтаксические модели в задачах обработки текста / Е.И. Яковчук, Е.А. Сидорова // Труды рабочего семинара «Наукоемкое программное обеспечение НПО-2011». Ершовская конференция по информатике. – Новосибирск: ИСИ СО РАН, 2011. – С.287-292.

[16] ***Апресян, Ю.Д.*** Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы / Ю.Д. Апресян, И.М. Богуславский, Б.Л. Иомдин, Л.Л. Иомдин, А.В. Санников, В.З. Санников, В.Г. Сизов, Л.Л. Цинман // Национальный корпус русского языка: 2003-2005. – М.: Индрик, 2005. – С.193-214.

[17] ***Blanco, X.*** Using NooJ for Multipurpose Analysis of Romance Languages Corpora / X. Blanco // Труды международной конференции «Корпусная лингвистика–2008». – СПб., 2008. – С.40–44.

[18] ***Кононенко, И.С.*** Жанровые аспекты классификации веб-сайтов / И.С. Кононенко, Е.А. Сидорова // Программная инженерия № 8. – 2015. – С.32–40.

[19] ***Загорулько, Ю.А.*** Автоматизация сбора онтологической информации об интернет-ресурсах для портала научных знаний / Ю.А. Загорулько // Известия Томского политехнического университета, 2008. - Т.312. - №5. - С. 114-119.

[20] ***Marino, J.*** Understanding SCA (Service Component Architecture) / J. Marino, M. Rowley. – Addison-Wesley Professional. – 2009. – 360 p.

[21] ***Garanina, N.*** A multi-agent text analysis based on ontology of subject domain / N. Garanina, E. Bodin, E. Sidorova // Proc. of 9th Ershov Informatics Conference (PSI'2014). – Novosibirsk: A.P. Ershov Institute of Informatics systems, 2014. – P.50-56.

[22] ***Garanina, N.*** A Distributed Approach to Coreference Resolution in Multiagent Text Analysis for Ontology Population / N. Garanina, E. Sidorova, I. Kononenko // Springer International Publishing AG 2018 A. K. Petrenko and A. Voronkov (Eds.): PSI 2017, LNCS 10742, 2018. - P.1–16.

[23] ***Garanina, N.O.*** Conflict resolution in multi-agent systems with typed relations for ontology population / N.O. Garanina, E.A. Sidorova, I.S. Anureev // Programming and Computer Software. – 2016. – Volume 42. – Issue 4. – P. 206–215.

[24] ***Серый, А.С.*** Идентификация объектов в задаче автоматической обработки документов / А.С. Серый, Е.А. Сидорова // Труды международной конференции Диалог'2011 «Компьютерная лингвистика и интеллектуальные технологии». – Вып.10(17). – М.: РГГУ, 2011. – С.580-590.

ONTOLOGY-BASED APPROACH TO MODELING THE PROCESS OF EXTRACTING INFORMATION FROM TEXT

E.A. Sidorova

A.P. Ershov Institute of Informatics Systems SB RAS, Novosibirsk, Russia
lsidorova@iis.nsk.su

Abstract

The article deals with models and methods of knowledge representation, focused on tasks of automatic text processing and information extraction. In the framework of our approach, information extraction is considered as a process of ontology population with information represented as instances of domain concepts. To describe this process three basic models are proposed. The model of the text representation defines the general scheme of text processing and provides the mapping of the received information on the text. The knowledge representation model includes a description of the subject vocabulary, genre models of the text and the models of facts, which allow modeling the processes of information extraction in terms of semantic classes of subject vocabulary and ontology of the subject domain. The attributive model of data representation ensures the preservation of information streams of data that arise in the process of extracting information, and allows the use of ontological methods for solving ambiguity problems and resolving the coreference. Thus, an original technique that allows users to design a text analysis system and simulate the information extraction based on the domain ontology is proposed.

Key words: *information extraction, model of text, domain vocabulary, model of fact, ontology population.*

Citation: *Sidorova EA.* Ontology-based approach to modeling the process of extracting information from text [In Russian]. *Ontology of designing*. 2018; 8(1): 134-151. - DOI: 10.18287/2223-9537-2018-8-1-134-151.

Acknowledgment

The work was financially supported by the Presidium of the SB RAS (Block 36.1 of the Comprehensive Program of the FNI SB RAS II.1) and the RFBR (grant No. 17-07-01600).

References

[1] ***Petasis G, Karkaletsis V, Paliouras G, Krithara A, Zavitsanos E***. Ontology Population and Enrichment: State of the Art. In Knowledge-driven multimedia information extraction and ontology evolution. LNAI 6050. Springer-Verlag Berlin; 2011: 134–166.

[2] ***Mel'chuk IA***. Opyt teorii lingvisticheskikh modelei «Smysl ↔ Tekst». Semantika. Sintaksis [Experience of Theory of Linguistic Models "Meaning ↔ Text". Semantics. Syntax] [In Russian]. – Moscow: Iazyki russkoi kul'tury Publ.; 1999.

[3] ***Narin'jani AS***. TEON-2: from Thesaurus to Ontology and backwards [In Russian]. DIALOG'2002: Proc.of the Int. Workshop. – Moscow: Nauka Publ. 2002; Vol.1: 307–313.

[4] ***Zagorulko YuA.*** Semantic technology for development of intelligent systems oriented on experts in subject domain [In Russian]. Ontology of designing. 2015; 1(15): 30-46.

[5] ***Dobrov BV, Ivanov VV, Lukashevich NV, Solov`ev VD***. Ontologii i tezaurusy`: modeli, instrumenty`, prilozheniia [In Russian]. – Moscow: Binom. Laboratoriia znanii`, Internet-universitet informatcionny`kh tekhnologii`; 2009.

[6] ***Rubashkin VSh***. Semanticheskii` komponent v sistemakh ponimaniia teksta [In Russian]. Artificial Intelligence: Proc. of the 10th Nat. Conf. with Int. Part CAI-2006. – Moscow: Fizmatlit; Vol.2: 455-463.

[7] ***Rozental` DE***. Upravlenie v russkom iazy`ke [In Russian]. – Moscow: Kniga publ.; 1986.

[8] ***Zhuravlyov AO***. Creating of Database of Governance Models for Russian Words [In Russian]. Artificial intelligence; 2013; 1: 91-97.

[9] ***Bol'shakova EI.*** Iazy`k leksiko-sintaksicheskikh shablonov LSPL: opy`t ispol`zovaniia i puti [In Russian]. Programmnye sistemy i instrumenty: Tematicheskii sbornik №15. – Moscow: MAX Press; 2014. - http://www.lspl.ru/articles/Paper_19_LSPL.pdf.

[10] ***Kovalev AI, Sidorova AE***. Tool for developing subject dictionaries based on lexical templates DigLex [In Russian]. Knowledge-Ontology-Theory (KONT-15): Proc. of Russian Conf. – Novosibirsk. Mathematics Institute of SB RAS; Vol.1; 2015: 123-130.

[11] ***Horoshevskii` VF***. OntosMiner: Semei`stvo sistem izvlecheniia informatcii iz mul`tiazy`chny`kh kollektcii` dokumentov [In Russian]. Artificial Intelligence: Proc. of the 9th Nat. Conf. with Int. Part CAI-2004. – Moscow: Fizmatlit; Vol.2: 573-581.

[12] ***Ermakov AE***. Automatical extraction of facts from texts of personal files: experience in anaphora resolution [In Russian]. Computational Linguistics and Intellectual Technologies: Proc. of the Int. Conf. "Dialogue". - Moscow: Nauka publ.; 2007: 131–135.

[13] ***Gershenzon LM, Nozhov IM, Pankratov DV***. Sistema izvlecheniia i poiska strukturirovannoi` informatcii iz bol`shikh tekstovy`kh massivov SMI. Arhitekturny`e i lingvisticheskie osobennosti [In Russian]. Computational Linguistics and Intellectual Technologies: Proc. of the Int. Conf. "Dialogue". - Moscow: Nauka Publ.; 2005: 97-101.

[14] ***Azarova IV, Grebenkov AS, Lando TM***. The context schema of predicate arguments for automatic expansion of a domain ontology [In Russian]. Computational Linguistics and Intellectual Technologies: Proc. of the Int. Conf. "Dialogue". - Moscow: RSUH; 2008; 7 (14): 11-16.

[15] ***Yakovchuk EI, Sidorova EA.*** Generalized semantic-syntactic models in text processing tasks [In Russian]. A.P. Ershov Informatics Conference (the PSI Conference Series, 11th edition). Science-Intensive Software SIS-PSI 2011 Satellite Workshop. – Novosibirsk: IIS SB RAS; 2011: 287-292.

[16] ***Apresian IuD, Boguslavskii IM, Iomdin BL, Iomdin LL, Sannikov AV, Sannikov VZ, Sizov VG, Tcinman LL***. Sintaksicheski i semanticheski annotirovannyi korpus russkogo iazyka: sovremennoe sostoianie i perspektivy [In Russian]. Russian National Corpus: 2003-2005. – Moscow: Indrik; 2005: 193-214.

[17] ***Blanco X.*** Using NooJ for Multipurpose Analysis of Romance Languages Corpora. Corpus Linguistics-2008: Proc. of the Int. Conf. – St. Petersburg; 2008: 40–44.

[18] ***Kononenko IS, Sidorova EA.*** Genre Aspects of Websites Classification [In Russian]. Software Engineering; 2015; N8: 32–40.

[19] ***Zagorulko YuA.*** Automation of ontologic information gathering about internet-resources for the scientific knowledge portal [In Russian]. Bulletin of the Tomsk Polytechnic University; 2008; 312(5): 114-119.

[20] ***Marino J, Rowley M.*** Understanding SCA (Service Component Architecture). – Addison-Wesley Professional; 2009.

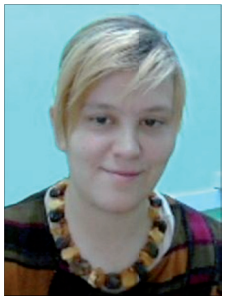
[21] ***Garanina N, Bodin E, Sidorova E.*** A multi-agent text analysis based on ontology of subject domain. Proc. of 9th Ershov Informatics Conference (PSI'2014). – Novosibirsk: A.P. Ershov Institute of Informatics systems; 2014: 50-56.

[22] ***Garanina N, Sidorova E, Kononenko I.*** A Distributed Approach to Coreference Resolution in Multiagent Text Analysis for Ontology Population. Springer International Publishing AG 2018 A. K. Petrenko and A. Voronkov (Eds.): PSI 2017. LNCS 10742; 2018: 1–16.

[23] ***Garanina NO, Sidorova EA, Anureev IS.*** Conflict resolution in multi-agent systems with typed relations for ontology population // Programming and Computer Software, July 2016. 2016; 42(4): 206–215.

[24] ***Seriy AS, Sidorova EA.*** Object identification in problem of automatic document processing [In Russian]. Computational Linguistics and Intellectual Technologies: Proc. of the Int. Conf. "Dialogue". - Moscow: RSUH; 2011; 10(17): 580-590.

Сведения об авторе



Сидорова Елена Анатольевна, 1977 г. рождения. Окончила Новосибирский государственный университет в 2000 г., к.ф.-м.н. (2006). Старший научный сотрудник лаборатории искусственного интеллекта Института систем информатики им. А.П. Ершова (Новосибирск) с 2014 г., член Российской ассоциации искусственного интеллекта. В списке научных трудов более 120 работ в области компьютерной лингвистики, мультиагентных систем, представления знаний и онтологического инжиниринга.

Elena Anatolievna Sidorova (b. 1977) has been senior researcher of the Laboratory of Artificial Intelligence at the A.P. Ershov Institute of Informatics Systems (Novosibirsk, Russia) since 2014, and was a leader of several projects in computational linguistics. She holds MSc degree from the Novosibirsk State University and PhD degree in Computer Science (2006). Dr. Sidorova has about 120 peer-reviewed publications in the field of NLP Systems, Multi-agent Systems, Knowledge Representation, and Ontology Engineering.