УДК 81`33, 004.8

Подход к извлечению информации из протоколов клинических испытаний на основе медицинской онтологии

Кононенко И.С. (Институт систем информатики СО РАН),
Сидорова Е.А. (Институт систем информатики СО РАН),
Боровикова О.И. (Институт систем информатики СО РАН)

В статье описан подход к организации процесса извлечения информации из протоколов клинических испытаний под управлением онтологии. Рассмотрены отдельные компоненты модели знаний, включая семантический словарь, жанровую модель текста, онтологию клинических испытаний, и приведены примеры извлечения конкретных ситуаций.

Ключевые слова: *извлечение информации, онтология предметной области, предметный словарь, жанр текста, модель факта, клинические испытания, медицинская онтология*.

1. Введение

В последнее время наблюдается резкий рост числа медицинских текстов, посвященных проводимым во всем мире клиническим исследованиям (КИ) лекарственных средств и медицинских технологий. Ежегодно в медицинской литературе появляется порядка 10 000 отчетов о новых завершенных рандомизированных клинических исследованиях [15]. Для получения представления о методике и результатах отдельных клинических исследований необходимо обращение к различным источникам, например, к базе медицинских публикаций MEDLINE. Информация о проводимых испытаниях фиксируется в виде протоколов, которые хранятся в специализированных базах, таких как международный реестр клинических исследований Национального института здоровья США www.clinicaltrials.gov, реестр Минздрава России www.grls.rosminzdrav.ru. Несмотря на свободный доступ к этим базам, поиск необходимой информации затруднен, поскольку отсутствует необходимая структуризация данных, а объемы выдачи исчисляются сотнями документов. С этой точки зрения представляют интерес системы, которые, работая с базой протоколов КИ, обеспечат: а) автоматическую индексацию текстов протоколов на основе онтологии, б) анализ и

структурирование описаний исследований в виде набора фактов, в) содержательный информационный поиск в базе на основе семантического анализа запроса пользователя.

Для обеспечения адекватного поиска необходимой информации в текстах статей, аннотаций и отчетов по КИ используются методы машинного обучения и автоматической обработки текста. В англоязычной литературе представлены работы этого направления, ориентированные на извлечение ключевых элементов КИ: [6-12,19]. Создаваемые системы опираются на существующие медицинские лексиконы и тезаурусы, такие как UMLS и MeSH (см., например, [10,19]). Конкретный набор извлекаемых ключевых элементов варьируется и во многом определяет применяемые методы обработки текста. Так, в [10] рассматривается задача извлечения информации из рефератов базы MEDLINE. Модули извлечения описаний пациентов, заболеваний, основных и сравнительных вмешательств используют правила, созданные вручную, в то время как модуль извлечения исходов основан на методах машинного обучения. Авторы мотивируют это тем, что заболевания и вмешательства описываются в тексте наименованиями, которые напрямую соответствуют концептам медицинского метатезауруса UMLS, описания пациентов представляются в виде шаблонов, включающих соответствующие концепты, в то время как описания исходов не имеют предсказуемой структуры и выходят за рамки именных групп, представляя собой большие фрагменты текста длиной от одного до восьми предложений.

Работа [7] посвящена поиску в текстах аннотаций КИ ключевых предложений, содержащих информацию о вмешательстве, параметрах исходов и участниках, с целью облегчить пользователю поиск релевантных фактов об экспериментальном дизайне КИ. Классификация осуществлена с помощью CRF-метода, для обучения использовался корпус структурированных рефератов. Описанная автоматическая разметка документов может далее использоваться не только для поиска и аннотирования, но и как первый шаг для идентификации и структурирования информации в рамках выбранных предложений. Именно такая двухступенчатая архитектура характеризует системы извлечения информации, описанные в [8,11,19]. В [8] целевая информация охватывает 23 информационных элемента (критерии отбора пациентов, размер выборки, параметры вмешательства, значения параметров исходов и т.п.), представленных в полнотекстовых публикациях о рандомизированных клинических исследованиях. Архитектура системы сочетает в себе текстовый SVM-классификатор, который осуществляет отбор предложений, предположительно содержащих искомую информацию, и поиск и извлечение целевых фрагментов текста, шаблоны которых описаны в виде простых регулярных выражений.

Особенностью содержания протоколов рандомизированных КИ является сопоставление двух (или более) вмешательств – экспериментального препарата и препарата сравнения – и соответствующих групп участников. Эта информация представляется в текстах посредством сравнительных и однородных конструкций, требующих более глубокого лингвистического анализа. В [11] для конструкций, представляющих сравнение видов применяемой терапии и сравнение сущностей, характерное для описаний параметров исходов, применяется семантический анализ. В [6] производится полный синтаксический анализ однородных конструкций, которые часто описывают сопоставляемые типы лечения. Полученные в результате синтаксические признаки используются статистическим классификатором.

Исследовательская работа, требующая анализа протоколов КИ, затруднена разнородностью формального представления информации в различных клинических областях. Вариантом решения этой проблемы является использование технологии Semantic Web для интеграции разнородных приложений на основе онтологии КИ, определяющей общие словарь и семантику [14]. Единое решение проблемы предложено в рамках проекта по созданию банка данных КИ (Trial Bank Project, http://rctbank.ucsf.edu/), которому посвящена серия публикаций: [8,15-18]. В [17] клинические исследования рассматриваются как разновидность научной деятельности человека. Соответствующим образом выстроена онтология OCRe, разработанная как OWL-онтология сущностей и отношений, представленных в протоколах КИ. Онтология не зависит от дизайна и клинической области исследования и ориентирована на интеллектуальную поддержку планирования и анализа КИ, включая поиск протоколов и оценку уже проведенных исследований по конкретной проблеме.

В данной работе предлагается подход к извлечению информации из протоколов КИ в рамках онтологического направления: информация словарей, семантико-синтаксических моделей и правил извлечения существенным образом опирается на структуру онтологии КИ. Второй особенностью предлагаемого подхода является ориентация анализа на специфику жанровой структуры протоколов, которые написаны на естественном языке, но подчиняются строгим требованиям не только к используемым наименованиям препаратов, но и к структуре изложения при описании процесса испытаний. Это дает возможность значительно ограничить область поиска информации путем использования условий на жанровый сегмент в правилах извлечения, благодаря чему высокая точность извлекаемых данных достигается без предварительного этапа классификации для поиска релевантных предложений. Структура извлекаемой информации, а также связь с жанровыми особенностями протоколов

задаются проблемной онтологией, которая фиксирует схему БД и способ ее наполнения данными, полученными в результате анализа текста.

2. Информационные потребности пользователя

Целевую информацию, отвечающую на основные вопросы доказательной медицины, принято представлять в виде фрейма PICO [9]: patient/problem (характеристики субъектов, отобранных для исследования/заболевание), intervention (вмешательство: диагностический тест, лекарственный препарат, терапевтическая процедура), comparison (с чем сравнивается исследуемое вмешательство: отсутствие вмешательства, другой препарат или процедура, плацебо), outcome (исход вмешательства – совокупность контролируемых параметров исходов). К базе MEDLINE обеспечен многоязыковой PICO-интерфейс [13]. Однако даже при формулировке запроса в формате PICO результаты поиска не удовлетворяют клиницистов ввиду огромных объемов выдаваемых ссылок и невозможности формулировки более детализированных информационных запросов.

Разрабатываемая информационная система ориентирована на поиск протоколов прошедших клинических испытаний, удовлетворяющих поисковым запросам различной сложности. Для русскоязычных исследователей актуален двуязычный поиск – как на русском, так и на английском языках.

Поисковые задачи:

- Поиск по ключевым терминам и тегам с учетом синонимов по всей структуре протокола;
- Поиск по сочетанию параметров/фактов;
- Поиск с учетом родо-видовых отношений и других отношений между сущностями (например, препарат – заболевание).

Аналитические задачи:

- Обработка количественных запросов по контролируемым параметрам (биостатистические показатели);
- Анализ успешности испытаний (доказанность основной статистической гипотезы).

3. Модель знаний

Знания о предметной области, используемые в предлагаемом подходе, опираются на модель предметной области, которая фиксирует понятия и отношения между ними в виде онтологии. Онтология КИ (см. Рис.1) содержит классы понятий *Клиническое испытание*, *Препарат*, *Заболевание*, *Группа*, *Цель*, *Результаты*, определяющие состав и условия

проведения испытаний и служащие для представления участников, объектов, целей и результатов КИ. Для отражения специальных знаний из области медицины, анатомии и фармакологии, неявно заданных в описаниях протоколов, в онтологию включены иерархии понятий *Лекарственных средств*, *Анатомических объектов* и *Химических веществ*, связанные с заболеваниями и между собой ассоциативными отношениями. В состав онтологии входят также понятия, относящиеся непосредственно к проведению и организации научной деятельности КИ, такие как *Организации*, *Персоны*, *Географические объекты*, *События*, *Документы*, *Методы исследования*.

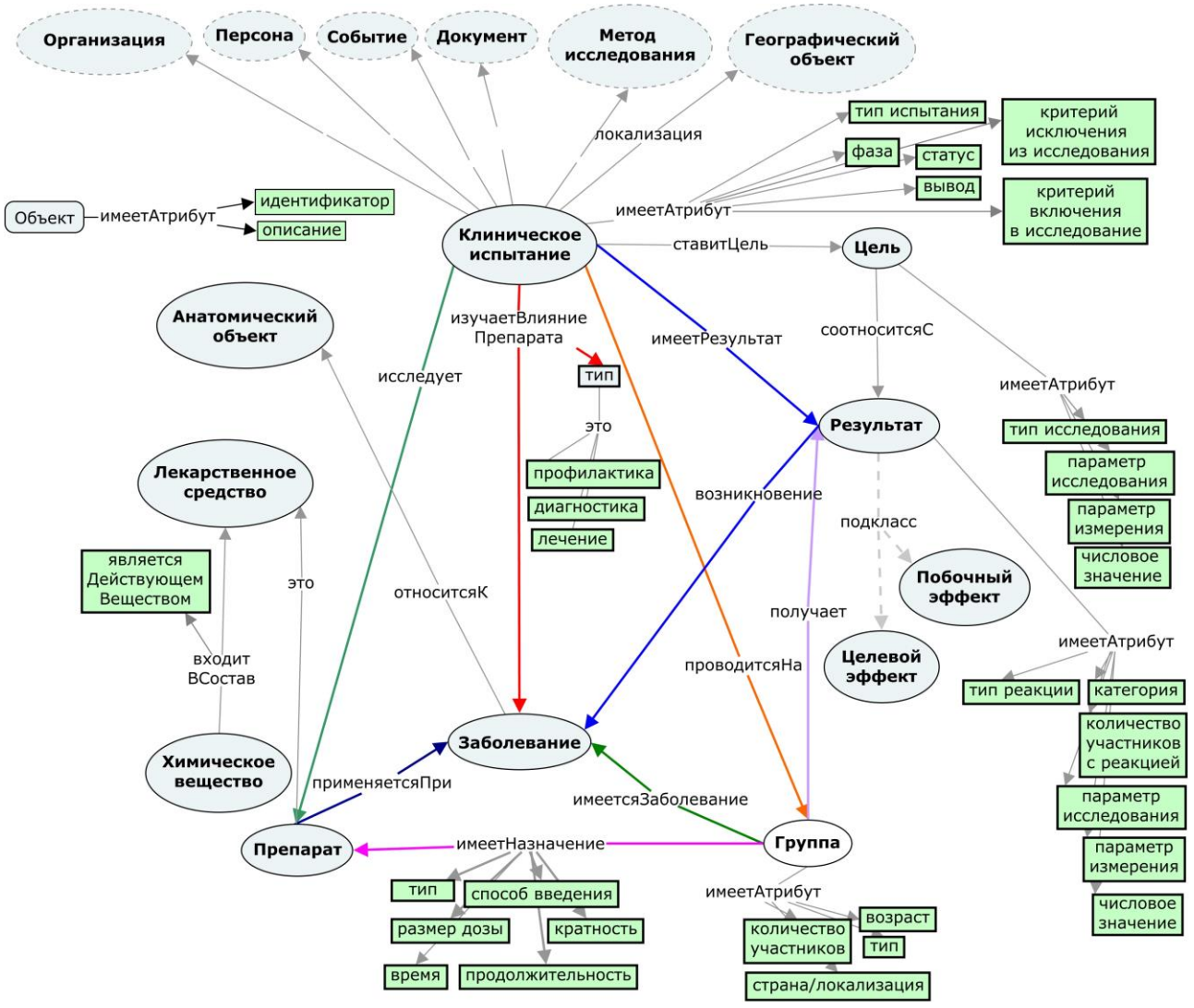

Рис. 1. Фрагмент онтологии предметной области "Клинические испытания".

Понятия онтологии КИ связаны между собой следующими основными отношениями:

«исследует» – связывает непосредственно данное КИ и исследуемый препарат;

«изучаетВлияниеПрепарата» – задает связь между КИ и заболеванием с указанием типа вмешательства: профилактика, диагностика, лечение, исследование качества жизни и т.п.

«проводитсяНа» – связывает КИ и группу участников, включенных в исследование;

«имеетНазначение» – определяет атрибутированную связь между группой участников-пациентов и исследуемым препаратом с заданием характеристик и условий приема препарата;

«имеетРезультат» – задает связь между КИ и полученными результатами (исходами) испытания;

«соотносится С» – позволяет задать связь между заданными целями и результатами, планируемыми и полученными в ходе испытаний.

На основе онтологии определяются характеристики информации для извлечения из доступных источников и способ (формат) ее представления для организации хранения и поиска.

Модель знаний о подъязыке предметной области представлена семантическими словарями (словарь предметной лексики, словари лексических шаблонов и семантико-синтаксических моделей управления), моделями фактов, описывающими способы выражения информации, принятые в рассматриваемой области знаний, а также знаниями об особенностях жанра рассматриваемых текстовых источников.

3.1. Текстовая коллекция

Корпус текстов содержит более 200 тыс. xml-документов, извлеченных из базы протоколов КИ, доступной на онлайн-ресурсе ClinicalTrials.gov. В базе представлены данные о клинических исследованиях широкого диапазона препаратов по различным показаниям. Формат и содержание протокола соответствуют принятым стандартам и положениям ICH GCP. В каждом протоколе представлены предопределенные форматом содержательные блоки (цель, задачи, дизайн, методология, статистические показатели исходов и др.), размеченные тегами и расположенные в строгой иерархической последовательности. Приведем фрагмент текста протокола, описывающий дизайн исследования:

```
<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>
```

Такая формальная схема разметки позволяет построить жанровую модель [1,5] рассматриваемых текстов в виде следующей упрощенной схемы.

```
PageGenre
   Block genre_segment
         Block genre_segment
         …
         Block genre_segment
   Block genre_segment
   …
```

Модель включает жанровые сегменты, маркированные жанровыми тегами, на основе которых извлекаются фрагменты текста для поиска той или иной информации.

Перечислим основные типы жанровых тегов рассматриваемых протоколов:

<brief_title>,<official_title>,<brief_summary>,<detailed_description> – описания типа, целей и содержания проводимых клинический испытаний;

<primary_outcome> – ожидаемые (целевые) результаты исследования;

<arm_group> – описание групп;

<intervention> – описание вмешательства;

<clinical_results>, <outcome> – описание результатов;

<reported_events>, <event> – побочные эффекты и т.д.

Жанровые особенности анализируемых текстов можно представить как совокупность следующих признаков.

Структурные особенности:

- Описание представлено иерархически организованными текстовыми блоками;
- Семантические единицы привязаны к структурным фрагментам текста.

Лексические особенности:

- Однозначность и единообразие терминов благодаря использованию номенклатурной лексики – наименований препаратов, заболеваний;
- Использование принятых аббревиатур и сокращений *ACAM200, US., PFU/ml, PRNT50*.

Грамматические особенности:

- Целевая информация представлена преимущественно редуцированными и полными именными группами, параметрическими и однородными конструкциями;
- Анафорические отсылки используются редко, представлены относительными местоимениями *which, who, that* и не выходят за рамки предложения.

Использование знаний о жанровых особенностях текста позволяет значительно ограничить разнообразие способов передачи информации, учитываемых в моделях фактов.

3.2. Словарь

Словарь системы создается путем обучения на представительном корпусе КИ, но ядро словаря составят термины тезауруса MeSH (https://www.nlm.nih.gov/mesh/), который в

версии 2016 г. содержит 27 883 дескрипторов, более 87 тыс. входных терминов-синонимов и 232 тыс. дополнительных концепт-записей – наименований конкретных химикатов, болезней и медикаментов.

Словарь состоит из основного словаря и словаря лексических шаблонов. В этих словарях фиксируется семантически значимая лексика, представляющая элементы целевой информации. Система семантических признаков в словарях основана на структуре онтологии клинических испытаний, отражая иерархию ее объектов и отношений. Объектные термины представлены преимущественно существительными (нарицательными и собственными именами), именными группами, лексическими конструкциями (аббревиатурами и более сложными буквенно-символьными конструкциями). С помощью семантических признаков объектные термины распределены по основным классам:

- Деятельность
    - Вмешательство (*intervention*)
    - Лечение (*therapeutics, acupuncture therapy, radiation treatment*)
    - Профилактика (*prophylaxis, preventive therapy, preventive procedure, vaccination, vaccinate*)
    - Диагностика/Обследование (*diagnostic procedure, diagnostic test, investigative techniques, blood chemical analysis*)
- Клиническое_испытание (*clinical trial, clinical study*)
- Препарат (*drug, organic chemicals, pharmaceutical, insulin lente, biological product, vaccine, herpesvirus vaccine, smallpox vaccine, ACAM2000, gD-Alum/MPL vaccine*)
- Болезнь
    - Вирусная_болезнь (*herpes henitalis, smallpox, hepatitis A*)
- Патологическое_состояние/признак/симптом (*asthenia, cyanosis, swelling, papule, pain, burning, itching, tingling, dysuria*)
- Состояние_здоровья (*healthy*)
- Анатомический объект
    - Система (*Cardiovascular System*)
    - Орган (*Heart, Miocardium*)
    - Локализация (*head, ear*)
- Участник (*participant, population*)
    - Персона (*patient, subject*)
    - Группа (*arm, cohort*)
- Организация
- Географический объект
    - Регион (*region, Europe*)
    - Страна (*Japan*)
    - Город (*city, New York, Moscow*)
- Временной объект

- Дата (*March 31, 2003*)
- Период (*from 10 January 2003 to 14 April 2003*)

Отдельный семантический класс Параметр составляют лексические единицы, описывающие параметрические характеристики объектов: доза, кратность, время и др. Объекты класса Участник описываются с помощью характеристик “этническая группа” (*african american, arab*), “пол” (*male, female, woman*), “возраст” (*age, aged, adult, adolescent, child, baby, 28 years, 50-59 years*). Группа и Клиническое испытание характеризуются признаком “тип” (*experimental, active comparator, placebo comparator*)**.** Назначение препарата характеризуют “доза” (*1.0x10-8th plaque-forming units/mL*), “время” (*on Day 0*), “способ” (*orally, parenteral, intramuscular*)**,** “кратность” (*single, twice*) и “продолжительность”.

Более детальная классификация терминов обусловлена такими онтологическими свойствами объектов, которые проявляются на уровне репрезентации в языковых конструкциях. Так, дополнительный семантический признак “колич” характеризует параметры, значения которых могут представляться нумеративной конструкцией. Признак “эталон” характеризует лексемы, представляющие стандартные оценки количественных параметров (*adult* vs. *24 to 34 years old, standard-dose, high dose* vs. *0.5 mL*). Специальные признаки выделяют элементы языковых конструкций параметрической семантики: “число” (*five*), “функция” (*more, equal, above, to*), “мера” (*milli-International Units, milliliter, microgram, mL*). Кроме того, отдельными признаками выделяются показатели временных отношений (*between <months 2 and 3>, preceeding, after*), однородности (*or, and*) и отрицание (*free <of>, without, not*).

Лексические признаки “тип”, “знач”, “имя” фиксируют особенности сочетаемости терминов в языковых конструкциях. Так, признаком “тип” выделены существительные-классификаторы, называющие объекты того или иного класса в общем виде (*vaccine*). Признак “имя” характеризует имена собственные (например, наименования препаратов *qHPV**,** Dryvax®, ACAM2000*).

Для извлечения дат и временных интервалов, сокращенных и стандартных наименований препаратов (*qHPV**,** ACAM2000*), значений параметров, представленных числовыми конструкциями (*<temperature> above 99.0°F, <dose:> 2.0x10-7th PFU/ml*), используется словарь лексических шаблонов. Шаблоны позволяют задать порядок следования элементов конструкций, описывающих наименования объектов, и учесть их написание с заглавной буквы, курсивом, латиницей, через дефис /тире или в кавычках. Так, типичная конструкция для дозы препарата представляется с помощью следующих шаблонов:

[кратное_число] = [число]**(_)**x**(_)**10**(_)-(_)** [цел_число] **(_)** (th)
[мера] =
plaque(_)(-)(_)forming unit…(_)(/)(_)ml
PFU(_)(/)(_)ml
[доза] = **[кратное_число] [мера]**
2.0x10-7th plaque-forming units/mL

Помимо семантически значимой лексики, словарь лексических шаблонов содержит класс Жанровой лексики. Это теги, содержащие слова и словосочетания (в том числе конструкции с подчерком), которые могут рассматриваться как индикаторы целевой информации. В процессе анализа используется структурированность блоков содержания с помощью разметки. Извлечение конкретных элементов целевой информации происходит в пределах выделенных индикаторами жанровых сегментов**.**

5. Поиск информации

При извлечении информации мы, помимо онтологии клинических испытаний, будем опираться на типичные информационные потребности пользователя-исследователя. Особый интерес представляют «содержательные» ситуации, описывающие проводимый эксперимент и его результаты. На текущий момент мы выделили четыре типа запросов:

1) Запросы, обеспечивающие поиск по характеристикам участников (пациентов) исследования, т.е. запросы на условия, применяемые к группам или когортам.
   Найти испытания, которые проводились над участниками
   - *с заболеванием* D,
   - *с расовой принадлежностью* R,
   - *возрастом до* A *лет*,
   - *живущих в стране* C,
   - *в которых применялась терапия типа* T.
2) Запросы, обеспечивающие поиск по особенностям применения препарата.
   Найти испытания, при которых применялся способ лечения препаратом P
   - *размер дозы меньше* X / *больше* X, *в интервале от* X1 *до* X2,
   - *вводится* Y *раз* / *однократно* / *многократно*,
   - *в течение времени* T / *меньше* T / *больше* T,
   - *способ доставки* W.
3) Запросы, обеспечивающие поиск по сочетанию ключевых элементов.
   Найти испытания, которые проводились
   - *для лечения заболевания* D/*для профилактики заболевания* D/*заболевания типа* TD,
   - *используя препарат* P/ *препарат типа* TP,
   - *с применением терапии типа* T.

4) Запросы, обеспечивающие поиск по результатам исследований.
   Найти успешные исследования / с серьезными побочными эффектами
   - *с применением терапии типа* T / *препарата* P,
   - *для лечения заболевания* D / *для профилактики заболевания* D / з*аболевании типа* TD,
   - *с наличием эффекта* X.

Таким образом, в соответствии с представленными типичными запросами, мы будем рассматривать следующие типы ситуаций: это, во-первых, описание участников испытаний и их деление по группам (*мужчины пожилого возраста, группа плацебо*), во-вторых, информация о препаратах, способах их применения, характере и особенностях проводимого лечения (*применение препарата в течение месяца 2 раза в день*), в-третьих, ситуации, характеризующие цели проводимых испытаний (*исследование безопасности дозы препарата*), и, наконец, описание полученных результатов (*положительный эффект был достигнут в 85% случаев*) и их соответствие поставленным целям.

В соответствии с представленными ситуациями сформирована схема универсальной ситуации, в которой имена классов выступают в качестве параметров поискового запроса:

Группа участвовала в *Испытании* с использованием *Препарата* для лечения/профилактики *Заболевания* с результатом *Результат* и побочным эффектом *Эффект*.

Найденные и распознанные ситуации можно представить в виде набора фактов, которые формально описываются экземплярами классов онтологии, значениями их атрибутов и связями. Для поиска и извлечения фактографической информации применяется технология анализа текста FATON [4], использующая ряд лингвистических ресурсов – терминологические словари, снабженные системой семантических признаков, а также лингвистическую модель предметной области КИ, содержащую набор *моделей фактов*, позволяющих в терминах семантических и грамматических признаков описывать способы выражения требуемой онтологической информации.

Каждая модель факта описывается схемой (правилом), которая включает набор аргументов структуры факта (arg1, arg2, …), их семантические/грамматические признаки, условия на семантико-синтаксическую сочетаемость характеристик аргументов, и набор объектов, который фиксирует структуру факта в онтологическом представлении. Рассмотрим несколько примеров и набор необходимых моделей для извлечения из них целевой информации.

4.1. Инициализация объектов. Как показано выше, система семантических признаков словаря формируется на основе онтологических сущностей, что позволяет инициализировать начальное формирование объектов непосредственно на основании словарных признаков.

Объект класса *Препарат* может быть представлен в тексте аппозитивной именной группой, в которой опорным словом является родовое слово или словокомплекс (тип), а имя примыкает к нему в постпозиции. Например,

Вакцина <Препарат, SemClass: тип> *"АСАМ2000"* <Препарат, SemClass: имя> [1]

извлекается с помощью модели:

Scheme Препарат3 : segment *Клауза* **(1)**
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 схеме термины должны иметь семантических класс *Препарат*, с учетом иерархии наследования признаков в словаре, а также первый термин должен обладать семантическим признаком *тип*, а второй – *имя*. На основе схемы создается 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 предпочтительным наименованием второго термина (Norm), заданным в тезаурусе для данного дескриптора (при наличии других входных терминов, синонимичных данному). Аналогичным образом могут извлекаться объекты *Заболеваний, Организаций* и т.п., если их названия присутствуют в словаре.

Инициализация объектов типа *Группа* возможна не только по названию, но и по присвоенному индексу, например, из фрагмента вида *“<group group_id="P5">”*.

Особо следует отметить случаи, 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, либо при последующей сборке ситуации (например, на основе семантической роли в ситуации).

Scheme Новый_объект : segment *Клауза* **(2)**
arg1: Term:: (SemClass: тип)
arg2: LexTerm::Именованный объект()
Condition PrePos(arg1,arg2), Contact(arg1,arg2)
⇨ Object :: Object (Тип: arg1.Class, Наименование: arg2.Name)

Появление таких объектов объясняется либо неполнотой базы знаний (например, при употреблении новых наименований препаратов, которые еще не зафиксированы в

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

онтологиях), либо наличием ошибок в тексте (в этом случае объект можно сопоставить с другими объектами того же типа, встречающимися в тексте рассматриваемого протокола).

4.2. Извлечение фактов. При поиске и выявлении характеристик объектов и их связей, как правило, требуется проверить сочетаемость семантических и/или грамматических признаков объектов. Для описания сочетаемости предикатных лексем разрабатывается словарь семантико-синтаксических конструкций (аналог моделей управления), который фиксирует семантические валентности предикатов, описывая их в терминах грамматических и семантических признаков актантов. Это позволяет проверять наличие управления в анализируемом фрагменте текста, т.е. согласованность семантических и синтаксических признаков предиката и актантов.

Рассмотрим примеры формирования с помощью моделей фактов фрагмента онтологии, описывающего характеристики объектов в рамках ситуации клинического испытания.

Извлечение характеристик объектов. Рассмотрим примеры схем, используемых для извлечения атрибутов объектов.

Scheme ТипКИсследования: genre_segment <*brief_title*> or <*official_title*> **(3)**
arg1: Term::Параметр(SemClass: цель_исследования)
arg2: Object::Препарат()
Condition PrePos (arg1,arg2), Contact(arg1,arg2), Упр(arg1,arg2)
⇨ Relation::изучаетВлияниеПрепарата (исследование: $this_CT, Препарат: arg2)
$obj1 = Object::Цель (тип: arg1.Name)
Relation::ставитЦель(исследование: $this_CT, цель: $obj1)

Данная схема позволяет извлекать информацию о цели проводимого исследования, зафиксированную в жанровых полях протокола <*brief_title*> или <*official_title*>. Так, из фрагмента *"Dose Study of ACAM2000 Smallpox Vaccine in Previously Vaccinated Adults …"* будет извлечен факт о том, что исследование посвящается изучению дозы вакцины от оспы.

Следующая схема позволит уточнить тип группы пациентов, принимающих участие в исследованиях.

Scheme ТипКогорты: genre_segment <*arm_group*> **(4)**
arg1: Object:: Группа(), genre_segment <*arm_group_label*>
arg2: Term::Параметр(), genre_segment <*arm_group_type*>
⇨ arg1: Группа (тип: arg2.Name)

Данная схема применяется для фрагментов вида:

<arm_group>
<arm_group_label>Group 5: Dryvax®</arm_group_label>
<arm_group_type>Active Comparator</arm_group_type>
</arm_group>

Создание отношений. Рассмотрим пример схемы построения отношения в соответствии с рассматриваемой ситуацией.

Scheme УсловияПримененияПрепарата: genre_segment <*group*> **(5)**
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)
⇨ Relation::имеетНазначение (группа: arg1, препарат: arg2, тип: «препарат»,
размер дозы: arg4.value, кратность: arg3.value, время: arg5.value)

Условия клинических испытаний для конкретной группы участников испытаний при назначении препарата описывается такими характеристиками, как наименование препарата, его дозировка, кратность применения и время приема. Данная информация в соответствии с принятым стандартом содержится строго в определенных жанровых фрагментах, однако в рамках фрагмента <*description*> описание параметров разворачивается в виде текста из одного-двух предложений, что требует применения более сложного лингвистического анализа (особенно в случаях комплексного применения препаратов).

Данная схема покрывает фрагменты вида:

<group group_id="P5">
<title>Dryvax® Vaccine</title>
<description> Participants received a single dose of Dryvax® smallpox vaccine, 1.0x10-8th plaque-forming units/mL on Day 0
</description>
</group>

В результате применения схемы к данному фрагменту текста будет создано описание ситуации терапевтического вмешательства для конкретной группы участников.

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

Разрешение кореференции объектов. Важной проблемой анализа текста является установление кореферентности объектов при их повторном упоминании. В общем случае онтологический подход позволяет разрешить кореференцию после основного анализа текста в процессе сравнения и идентификации объектов (относительно онтологии). Эквивалентными с точки зрения онтологии считаются объекты с непротиворечивыми классами и наборами атрибутов [2].

Особенности протоколов КИ, содержащих однозначную номенклатурную лексику, распределенную по разным структурным блокам, упрощают процедуру поиска эквивалентных с точки зрения онтологии объектов.

Заключение

В статье описан подход к организации процесса извлечения информации под управлением онтологии. Рассмотрены отдельные компоненты системы и приведены примеры извлечения конкретных ситуаций, описывающих клинические испытания.

В нашей лаборатории создан ряд инструментов, поддерживающих все этапы разработки и эксплуатации систем извлечения информации с опорой на онтологию [3]. Для разработки информационной базы клинических испытаний используются следующие инструменты: а) технология построения предметных словарей KLAN, поддерживающая методы машинного обучения, тематической и жанровой классификации, морфологического и поверхностно-синтаксического анализа текстов и обеспечивающая эксперта-лингвиста широким набором инструментов для отладки словаря; б) технология построения лексических шаблонов DigLex, обеспечивающая поиск в тексте несловарных конструкций, таких как сокращения, буквенно-числовые обозначения препаратов, химические названия веществ и т.п.; в) система жанровой сегментации текстов; г) система фактографического анализа текстов FATON, которая реализует обработку текста на основе схем фактов и обеспечивает пополнение БД системы; д) технология построения портала знаний, обеспечивающая доступ пользователей к информационному наполнению базы данных, содержательный поиск и навигацию на основе онтологии.

Планируется апробировать предложенный подход на публичной базе английских текстов клинических испытаний, представленных на сайте *ClinicalTrials.gov*, и в дальнейшем направить усилия на создание аналогичной системы для русскоязычного контента.

Список литературы

1. Кононенко И. С., Сидорова Е. А. Жанровые аспекты классификации веб-сайтов // Программная инженерия. 2015. № 8. С. 32–40.
2. Серый А.С., Сидорова Е.А. Поиск референциальных отношений между информационными объектами в процессе автоматического анализа документов // Труды XIV Всероссийской научной конференции RCDL-2012 Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Переславль-Залесский, 2012. С. 206-212.
3. Сидорова Е.А. Разработка лингвистического обеспечения информационных систем на основе онтологических моделей знаний // Известия Томского политехнического университета. 2013. Т. 322. № 5. С. 143-147.
4. Сидорова Е.А. Фактографический анализ текста в контексте интеллектуальных информационных систем // Информационные и математические технологии в науке и

управлении: тр. XVIII Байкальской Всероссийской конференции. Иркутск: Институт систем энергетики им Л.А. Мелентьева СО РАН, 2013. Т.3. С. 79-85.

5. Сидорова Е.А., Кононенко И.С. Представление жанровой структуры документов и ее использование в задачах обработки текста // Труды Седьмой Международной конференции памяти академика А.П. Ершова "Перспективы систем информатики". Рабочий семинар «Наукоемкое программное обеспечение». Новосибирск: Сибирское Научное Издательство, 2009. С. 248-254.
6. Chung G.Y. Towards identifying intervention arms in randomized controlled trials: extracting coordinating constructions // Journal of Biomedical Informatics. Vol. 42. 2009. P. 790–800.
7. Chung G.Y., Coiera E. A study of structured clinical abstracts and the semantic classification of sentences // Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. 2007. P. 121–128.
8. De Bruijn B., Carini S., Kiritchenko S., Martin J., Sim I. Automated information extraction of key trial design elements from clinical trial publications // Proceedings of AMIA Annual Symposium. 2008. P. 141-155.
9. Demner-Fushman D., Lin J. Answering clinical questions with knowledge-based and statistical techniques. Computatinal Linguistics. Vol.33 (1). 2007. P. 63-103.
10. Demner-Fushman D, Lin J. Knowledge Extraction for Clinical Question Answering: Preliminary Results // Proceedings of AAAI Workshop on Question Answering in Restricted Domains. 2005. P. 1–9.
11. Fiszman M, Demner-Fushman D, Lang FM, Goetz P, Rindflesch T. Interpreting comparative constructions in biomedical text // Proceedings of the BioNLP workshop, association for computational linguistics. 2007. P. 137–44.
12. Ke-Chun Huang, I-Jen Chiang, Furen Xiao, et al. PICO element detection in medical text without metadata: Are first sentences enough? // Journal of Biomedical Informatics. Vol.46. 2013. P. 940-946.
13. PICO Linguist. [Electronic resource]. URL: http://babelmesh.nlm.nih.gov/pico.php (Accessed: 9/05/2017).
14. Ravi D. Shankar, Susana B. Martins, MD, Martin O'Connor, David B. Parrish, Amar K. Das. An Ontology-based Architecture for Integration of Clinical Trials Management Applications. //Proceedings of AMIA Annual Symposium. 2007. P. 661-665.
15. Sim I. The Trial Bank Project. [Electronic resource]. URL: http://grantome.com/grant/NIH/R01-LM006780-10 (Accessed: 9/05/2017).
16. Sim I., Olasov B., Carini S. An ontology of randomized controlled trials for evidence-based practice: content specification and evaluation using the competency decomposition method. Journal of Biomedical Informatics. Vol.37. 2004. P. 108-119.

17. Sim I., Tu Samson W., Carini S. et al. The Ontology of Clinical Research (OCRe): an informatics foundation for the science of clinical research // Journal of Biomedical Informatics. Vol.52. 2014. P. 78-91.
18. Tu Samson W., Peleg M., Carini S.et al. A practical method for transforming free-text eligibility criteria into computable criteria // Journal of Biomedical Informatics. Vol.44, 2011. P. 239-250.
19. Xu R, Garten Y, Supekar KS, Das AK, Altman RB, Garber AM. Extracting subject demographic information from abstracts of randomized clinical trial reports // Studies in Health Technology and Informatics. Vol.129. 2007. P. 550–554.