ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ № 3(12), 2012, с. 57–70

УДК 004.89:004.912

Е. П. Куршев, Е. А. Сулейманова, И. В. Трофимов

Роль знаний в системах извлечения информации из текстов

Аннотация. В работе рассматриваются стадии процесса извлечения информации из текстов на базе правил. Затрагивается проблема применения формализованных знаний о предметной области на стадиях синтактико-семантического анализа и собственно извлечения. Подчеркивается значимость привнесения «фоновых» знаний в результат извлечения; рассматривается способ реализации этого функционала.

Ключевые слова и фразы: извлечение информации, «фоновые знания», семантический анализ.

Введение

Быстрый рост сети Интернет и увеличение объемов электронного документооборота делают технологии автоматической обработки текстов все более востребованными. Среди них значительный интерес представляют *методы извлечения информации*, позволяющие структурировать содержащуюся в текстах информацию, что делает возможным дальнейшую ее обработку традиционными алгоритмическими методами.

Извлечение информации находит применение во многих областях. Так, например, в финансовой сфере оно используется для выявления в новостных потоках упоминаний о сделках слияния/поглощения. В кадровом бизнесе такие технологии применяются для автоматической обработки резюме соискателей и объявленных вакансий. Маркетинговые отделы используют извлечение информации для анализа отзывов об имеющейся на рынке продукции. Все это разные по сложности задачи, требующие для своего решения различных ресурсов и подходов.

Работа выполнена при финансовой поддержке Министерства образования и науки Российской Федерации (госконтракт № 07.514.11.4109).

Сложность задачи извлечения зависит от многих факторов:

- сложность текста
  - сложность языка, на котором написан текст;
  - качество текста (грамотность автора);
  - наличие или отсутствие частичной структуры документа;
  - наличие априорных знаний о содержании текстов;
  - присутствие или отсутствие стилевой/жанровой однородности текстов;
- сложность извлекаемой (целевой) информации
  - множественность способов изложения искомых фактов в текстах;
  - вид извлекаемой информации (поиск взаимосвязей сложнее поиска отдельных сущностей).

Новостные сообщения, например, отличаются высоким качеством текста, но отсутствует структура документа, и стилевая однородность также может отсутствовать (если сравнить тексты, публикуемые новостными агентствами и обзорно-аналитическими новостными СМИ). С другой стороны, тексты резюме, как правило, имеют какую-то структуру, их содержание априорно известно и ограничено, но качество текста можно охарактеризовать как среднее. Наконец, при анализе произвольных интернет-источников (форумов, блогов) приходится иметь дело с текстами низкого качества. Это, вероятно, наиболее сложные для анализа тексты.

Современные методы извлечения информации довольно поверхностны и опираются, как правило, на лингвистические характеристики текста, лексику и ограниченные предметные знания. Это приводит к их неспособности решать на приемлемом уровне качества сложные задачи извлечения. В данной статье рассматривается, на каких этапах анализа и каким образом могли бы использоваться предметные знания для повышения качества анализа.

1. Современные методы извлечения информации

Современные методы извлечения информации из текста делятся на два больших класса: статистические и на базе правил [1]. Статистические методы неплохо справляются с задачами извлечения, когда целевая информации представлена последовательно, например, извлечение компонентов адреса или библиографической записи. Методы на базе правил более универсальны и могут применяться

для решения любых задач извлечения. Их, в свою очередь, можно разделить на инженерные и обучаемые.

Основой и инженерных, и обучаемых методов являются правила извлечения информации, описывающие, как найти целевую информацию в тексте [2, 3]. Инженерный подход предполагает разработку таких правил людьми (лингвистами, специалистами по предметной области). Получаемые в результате правила извлечения обычно получаются довольно общими и компактными, что положительно сказывается на производительности системы извлечения. Главный недостаток подхода — необходимость привлечения высококвалифицированных специалистов.

Методы на базе машинного обучения порождают правила извлечения автоматически, используя размеченный текстовый корпус. Преимуществами такого подхода являются отсутствие необходимости привлечения высококлассных специалистов-предметников и более простая адаптация к новым, не учтенным ранее, способам выражения целевой информации (нужно переобучиться). Существует мнение, что создание систем извлечения информации на базе машинного обучения менее трудоемко, однако это спорный вопрос, так как для достижения хороших результатов требуется разметка довольно большого текстового корпуса. И чем менее формализованы тексты, шире спектр целевой информации и сложнее язык, тем большего размера нужен корпус.

Большинство известных нам подходов к извлечению информации опираются на скудные предметные знания. Как правило, их использование ограничивается определением семантического класса слова (словосочетания). Даже среди тех систем, которые опираются на онтологические ресурсы [4], далеко не все выходят за рамки этого ограничения. В то же время эти подходы, видимо, достигли предела своих возможностей, и дальнейшее их развитие без более широкого использования предметных знаний невозможно.

Вместе с тем к самой задаче извлечения информации предъявляют все более высокие требования. Изначально задача предполагала автоматическую разметку в тексте целевой информации или ее извлечение в несложный и небольшой фрейм. Сейчас в ее рамках пытаются решать подзадачи выявления невыраженного явно смысла, унификации и отождествления результатов извлечения, привязки извлекаемой информации к элементам сложных ресурсов

знаний, например онтологий. Для качественного их решения необходимы знания общего характера об устройстве мира, не содержащиеся в тексте, но которые человек приобретает по мере накопления жизненного опыта и использует при чтении текстов. Также нужны механизмы использования этих знаний в процессе анализа документа.

Рассмотрим подробнее, на каких стадиях анализа текста и каким образом могут использоваться предметные знания. Приведенные соображения основываются на нашем опыте работы с русским языком, хотя в большинстве своем они носят общий характер и применимы при анализе других языков.

2. Стадии анализа текста

Процесс анализа текста в системах извлечения информации принято делить на несколько стадий, каждая из которых решает одну определенную задачу. Полноценная система извлечения информации включает следующие шаги.

- **Подготовка текста.** Сюда входит разбор формата документа (извлечение текста, информации о структуре и стилевом оформлении текста), определение языка и кодировки текста, перекодировка (при необходимости), простая корректировка опечаток или ошибок оптического распознавания.
- **Графематический анализ** — преобразование текста в последовательность токенов (слов и других символьных последовательностей, образующих законченный языковый объект) и их классификация. Поиск возможных границ предложений.
- **Морфологический анализ** — определение частей речи и морфологических атрибутов у слов текста.
- **Постморфологический анализ** — имеет дело с частными случаями морфологического анализа, не укладывающимися в общую систему. Например, морфологический анализ для фамилий нецелесообразно решать словарными методами (потенциально множество фамилий бесконечно) или общими эвристическими методами (высока вероятность ошибки). Для них создаются *специальные* эвристические методы. К фазе постморфологического анализа может относиться также снятие морфологической омонимии.

- **Полный или частичный синтактико-семантической анализ** — выявление синтаксических связей в предложениях, их синтаксическая и семантическая классификация, а также семантическая классификация участников синтаксических связей.
- **Разрешение кореферентности** — установление референциального тождества разных упоминаний об одной сущности.
- **Извлечение информации** — поиск в тексте целевой информации и преобразование ее в структурированную форму. Для флективных языков на этом этапе решается также задача нормализации.
- **Унификация и отождествление извлеченной информации. Выявление информации, не представленной в явной форме** — проецирование извлеченной информации в единую концептуальную модель; окончательный переход от текстового представления к смысловому.

3. Знания и синтаксический анализ

Предметные знания играют ключевую роль, начиная с синтаксического анализа. В русском языке опора на знания необходима уже на стадии микросинтаксиса — определения границ именных групп и синтаксических связей внутри группы. Так, если адъективные связи можно достаточно уверенно определять на основе только формальных морфологических атрибутов, то в случае с генитивными связями возникают неоднозначности.

Генитивы могут образовывать линейные цепочки различной длины:

- *«директор компании»*;
- *«заместитель директора компании»*;
- *«по мнению заместителя директора компании»*.

При этом нет надежного «локального» лингвистического критерия, позволяющего определить конец цепочки. Например, в предложении *«В вестибюле гостиницы слесаря не оказалось.»* морфологические атрибуты не препятствуют установлению генитивной связи между гостиницей и слесарем (с семантикой «гостиница принадлежит слесарю»), и только макросинтаксический анализ может помочь отвергнуть этот вариант, так как предложение получится

«неполным». Использование знаний могло бы помочь нам разобраться с этой проблемой и подсказать, что гостиницы обычно не принадлежат слесарям.

Морфологическая омонимия еще более усугубляет ситуацию с генитивом. Рассмотрим предложение *«Депешу президента атташе получил только через час»*. Для несклоняемого «атташе» морфологический анализ предложит все варианты падежей, в том числе и родительный. Поэтому формально между президентом и атташе может существовать генитивная связь. Однако среди возможных семантик генитивной связи нет такой, которая могла бы иметь место в реальном мире с данной парой аргументов (президент, атташе).

Наконец, генитивы не всегда образуют цепочку; могут быть дистантные связи. Поэтому, кроме определения границ синтаксической группы, неоднозначность проявляется и в вопросе, между какими словами устанавливать связь? В качестве примера можно привести фразу *«отдел закупок крупной компании»*. Здесь генитивная связь существует между парами «отдел, закупок» и «отдел, компании». При выборе пары кандидатов для связывания недостаточно формальных критериев (правила проективности, падежей и т.д.) — вариант «закупок, компании» формально допустим, но семантически неприемлем. Для разрешения этой проблемы знания являются необходимыми.

Мы рассмотрели только примеры, касающиеся наиболее проблемного отношения русского микросинтаксиса — генитива. Но, разумеется, этим сфера применения знаний в синтактико-семантическом анализе не ограничивается.

4. О семантическом анализе

Семантика как раздел лингвистики изучает отношения между выражениями естественного языка и действительным (или воображаемым) миром [5]. При этом в качестве значения языкового выражения могут выступать различные вещи. Рассмотрим их подробнее, опираясь на диаграмму (рис. 1), прообразом для которой стал известный треугольник Фреге [6].

Современная теория *лингвистической* семантики определяет денотат как класс сущностей, по отношению к которым может быть использовано данное языковое выражение (слово). В свою очередь, сигнификат — это определяющие признаки сущности, служащие условием применимости языкового выражения. Те признаки, при

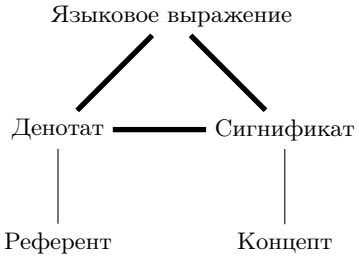

Рис. 1. Языковое выражение и его значения

отсутствии которых данное языковое выражение к данной сущности не приложимо вообще [5]. В традиционной «слабой» семантике значением языкового выражения считается его сигнификат.

С позиции «сильной» семантики значением языкового выражения считается референт — та конкретная сущность реального или воображаемого мира, к которой идет отсылка (при помощи языкового выражения) в данном конкретном высказывании. Референт соотносится с денотатом как элемент с множеством, то есть референт является экземпляром денотата (как класса сущностей).

В задаче *извлечения информации* интерес представляет сильносемантическая интерпретация языкового выражения, так как нас обычно интересуют конкретные факты о конкретных объектах действительности. Рассмотрим, например, предложение «Президент России отправился в США, а американский президент — в Россию». Несмотря на то, что сигнификат слова «президент» один и тот же, мы здесь имеем дело с двумя отдельными фактами о двух различных президентах (референты различны).

В процессе *синтаксического анализа* для разрешения неоднозначностей достаточно слабо-семантической интерпретации. В приведенном выше примере про президента и атташе знания концептуального уровня о том, что должности президента при атташе не бывает, «запрещают» установление синтаксической связи. С другой стороны, должность секретаря при президенте бывает, поэтому в предложении «Мы встретили секретаря президента» генитивная синтаксическая связь между секретарем и президентом возможна.

Необходимо учитывать, что не только референты, но и сигнификаты одного и того же слова могут быть различны (полисемия,

омонимия). В предложении «Президент компании отправился в круиз» слово «президент» ссылается на некоторое лицо, о котором мы ничего не знаем, кроме того, что он занимает должность президента. Вместе с тем словом «президент» также обозначается соответствующая должность. В предложении «Должность президента вакантна» подразумевается именно этот смысл. Таким образом, мы имеем два сигнификата слова «президент»:

- с одной стороны, это руководящая или представительская должность,
- с другой — занимающее эту должность лицо.

Для извлечения информации правильное определение семантического класса извлекаемых сущностей является необходимым условием.

Рассмотрим теперь различие между сигнификатом (языковым значением) и концептом (понятием). Это различие хорошо описано в энциклопедии Кругосвет [5]. *«Если понятие — это полное (на данном уровне познания) отражение в сознании признаков некоторой категории объектов или явлений, то языковое значение фиксирует лишь их различительные признаки. Так, в значение слова река входят такие «дифференциальные признаки» понятия о реке, как «водоем», «незамкнутый», «естественного происхождения», «достаточно большого размера», по которым объект, именуемый рекой, отличается от объектов, именуемых канавой, морем, прудом, озером, ручьем. Понятие же о реке включает, помимо данных, и другие признаки, например «питающийся за счет поверхностного и подземного стока своего бассейна». Можно сказать, что значению слова соответствует «наивное», обиходное понятие о предмете (в отличие от научного)».*

Априорные знания (как в форме сигнификата, так и в форме концепта) позволяют представить в явном виде те сведения (факты), которые в тексте в развернутой форме не присутствуют, но скрыты в самом понятии. Эти сведения в дальнейшем могут использоваться алгоритмами анализа и поиска в извлеченной информации. Из прикладных соображений кажется разумным иметь дело именно с концептами, так как для поиска и анализа могут потребоваться не только дифференциальные признаки, но и другие. Для компьютерного моделирования концептуализаций и работающих с ними процедур вывода разработаны подходы, известные как онтологический инжиниринг.

5. Знания и извлечение информации

Извлечение информации можно рассматривать как инструмент для создания фактографического информационного ресурса, в котором затем будут выполняться поисково-аналитические процедуры, реализующие полезную бизнес-логику. Чем больше информации будет извлечено в информационный ресурс, тем шире будут возможности и выше качественные характеристики поисковых и аналитических процедур бизнес-логики.

Что в данном случае понимается под словами «больше информации»? Речь идет о степени структурированности информации (гранулярности) и степени ее представления в явном виде.

Использование в данном контексте понятия гранулярности обусловлено тем, что один и тот же фрагмент текста может быть проинтерпретирован с различной степенью глубины. Например, фразу *«заместитель директора»* мы можем рассматривать как атомарно (название должности), так и структурно (отношение между двумя должностями — должностью заместителя и должностью директора). Таким образом, гранулярность определяет, насколько мелкие структурные единицы мы получим в результате интерпретации информации, содержащейся в тексте в явном виде.

В то же время значительная часть информации содержится в тексте в неявном виде. Этот факт отмечался в ранних работах по автоматическому анализу естественного языка, однако не нашел отражения в традиционных методах извлечения информации. Шенк пишет [7]: *«Многое из содержания фраз естественного языка остается невыраженным явно (например, «жидкость» в фразе «Хотите пить?» или «деньги» — в «Я купил книгу»)»*. Сталкиваясь с этой прагматической составляющей коммуникации, человек использует «фоновые» знания для восстановления (при необходимости) полного смысла фразы. Чтобы программа имела возможность, проанализировав текст, отвечать на те же вопросы, на которые, прочитав этот текст, может ответить человек, необходимо предоставить программе эти «фоновые» знания, причем в достаточно формализованном виде, чтобы алгоритмы могли их использовать.

Фоновые знания могут проявляться при интерпретации как отдельного слова, так и фразы. Рассмотрим несколько примеров.

- Словом *«пароходство»* кодируется не только семантический класс «организация», но также набор атрибутивной информации о том, что данная организация занимается перевозками, причем при помощи водных транспортных средств.
- Слово *«министр»* предполагает существование организации вида «министерство», которую этот министр возглавляет.
- Фраза *«акционеры компании»* кодирует не только отношение между двумя сущностями (акционерами и компанией), но и содержит сведения о том, что компания по организационно-правовой форме является акционерным обществом.
- Во фразе *«оптовая компания»* признак *оптовый* косвенно указывает на род деятельности компании — *торговля*.

Глубокая интерпретация перечисленных слов и фраз, вскрывающая фоновые знания, позволила бы традиционными алгоритмическими средствами (например, используя SQL) реализовать такие бизнес-функции, как подсчитать количество упоминаний транспортных предприятий в текстовой коллекции, найти события заданного типа с участием акционерных обществ или найти названия торговых предприятий.

В своем подходе к извлечению информации мы используем различные механизмы хранения и «вскрытия» фоновых знаний. Для представления предметных знаний (а также результатов извлечения информации) нами используется модель, близкая к фреймовой [8, 9]. Каждый концепт описывается фреймом с индивидуальной слотовой структурой, определяющей признаки, которыми могут характеризоваться объекты, соответствующие данному понятию. В этой модели фоновые знания представляются в форме значений по умолчанию и ограничений на значения слотов фрейма. Первые служат для представления информации о конкретном значении признака, характеризующего данное понятие, а вторые — для ограничения спектра возможных значений признака. Например, для понятия «банк» в слоте «сфера деятельности» мы можем указать значение по умолчанию «финансовая», в то время как у понятия «президент» в слоте «возглавляет что» можно указать лишь ограничение «или организация, или геополитическое образование», а конкретное значение данный слот может принять только на основании информации, содержащейся в анализируемом тексте.

Возвращаясь к примерам с пароходством и акционерами, покажем, каким образом осуществляется отражение фоновых знаний в

результатах извлечения. Мы не будем здесь рассматривать проблемы неоднозначности (полисемии, омонимии, неоднозначности синтаксических связей), предполагая, что данная проблема разрешена. То есть уже известно, какому концепту соответствует данный фрагмент текста, а синтаксические связи построены правильно.

Сначала рассмотрим случай, когда фоновые знания являются составляющей концепта и не зависят от контекста его употребления — *пароходство*. В нашей модели предметной области пароходство, как всякая организация, характеризуется слотом «род деятельности». Для пароходства этот слот заполняется значением по умолчанию «перевозки». Кроме того, этот концепт имеет слот «вид транспорта», заполненный по умолчанию значением «водный». Благодаря этому, представляющие интерес референтные употребления [10] *пароходства* в процессе извлечения будут преобразованы в объекты типа «организация», у которых указанные слоты заполнятся автоматически.

В случае с «министром» фоновые знания также вскрываются бесконтекстно, но используется иной механизм получения их в результатах извлечения. В данном случае нам нужно не охарактеризовать министра дополнительной информацией (заполнить его слот), а «реконструировать» не упоминаемый явно объект «министерство» и отношение между министром и министерством. Для этих целей мы используем формальный язык (язык трансформаций извлеченной информации) [9], в котором кодируются фоновые знания. На этом языке описываются правила преобразования структурированной информации, полученной в ходе извлечения.

Теперь рассмотрим случаи, когда имплицитная информация восстанавливается благодаря контексту, в котором употребляется понятие — *акционеры компании*. В силу наличия синтаксической связи между словами *акционеры* и *компании*, построенные в результате извлечения объекты оказываются связаны предметным отношением (интерпретацией синтаксической связи) — «роль по отношению к». В то же время в модели предметных знаний содержится информация о том, что объект типа «акционер» может быть связан этим отношением только с объектом типа «акционерное общество». Это позволит заполнить слот «организационно-правовая форма» компании значением «акционерное общество».

Другой случай связан с интерпретацией признаковой информации — *оптовая компания*. Признак «оптовый» является значением слота «вид торговли» у концепта «торговля». В свою очередь, «торговля» подчинена (в таксономической иерархии) концепту «род деятельности». Это позволяет сделать вывод о том, что род деятельности упомянутой компании — торговля. Такой вывод мы реализуем средствами языка трансформаций.

6. Заключение

Мы показали, что предметные знания имеют огромное значение для разрешения неоднозначностей синтаксиса и для полной интерпретации извлекаемой информации (привнесения фоновых знаний в результат). Не менее значимы они и для решения задач разрешения кореферентности и отождествления извлеченной информации. Приведенные в статье примеры демонстрируют, какого рода знания востребованы для решения задачи извлечения информации из текста, однако форма их представления и использования является предметом дальнейших исследований.

Список литературы

[1] Sarawagi S. *Information Extraction* // Foundations and Trends in Databases, 2008. **Vol. 1**, no. 3, p. 261–377 ↑1

[2] Appelt D. E. The Common Pattern Specification Language: Technical report. Menlo Park : SRI International, Artificial Intelligence Center, 1996. ↑1

[3] Кормалев Д. А., Куршев Е. П. *Развитие языка правил извлечения информации в системе ИСИДА-Т* // Тр. междунар. конф. "Программные системы: теория и приложения". — ИПС им. А.К. Айламазяна РАН, Переславль-Залесский, 2006, с. 365–377 ↑1

[4] Wimalasuriya D.C. *Ontology-based information extraction: An introduction and a survey of current approaches* // Journal of Information Science, 2010, June. **Vol. 36**, no. 3, p. 306–323 ↑1

[5] *СЕМАНТИКА* // Энциклопедия Кругосвет, http://www.krugosvet.ru/enc/gumanitarnye_nauki/lingvistika/SEMANTIKA.html ↑4, 4

[6] Степанов Ю. С. Семиотика. М. : Наука, 1971. ↑4

[7] Шенк Р. Обработка концептуальной информации. Пер. с англ. М. : Энергия, 1980. ↑5

[8] Александровский Д. А., Кормалев Д. А., Куршев Е. П., Сулейманова Е. А., Трофимов И. В. *Модель и реализация ресурса знаний в системе извлечения информации из текста* // Тр. Одиннадцатой нац. конф. по искусственному интеллекту с междунар. участием КИИ-2008. — Дубна, Россия, 2008. Т. **2**, с. 201–209 ↑5

[9] Кормалев Д. А., Куршев Е. П., Сулейманова Е. А., Трофимов И. В. *Извлечение информации из текста в системе ИСИДА-Т* // Труды XI Всероссийской научной конференции RCDL'2009. — Петрозаводск, 2009, с. 247–253 ↑5

[10] Арутюнова Н. Д. Предложение и его смысл: Логико-семантические проблемы. М. : Наука, 1976. ↑5

Рекомендовал к публикации *к.т.н. Е. П. Куршев*

Об авторах:

Евгений Петрович Куршев

Руководитель Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна РАН.

e-mail: epk@epk.botik.ru

Елена Анатольевна Сулейманова

Научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна РАН, одна из разработчиков технологии построения систем извлечения информации.

e-mail: yes@helen.botik.ru

Игорь Владимирович Трофимов

Старший научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна РАН, специалист по технологиям автоматической обработки естественного языка, извлечения информации, автоматического планирования.

e-mail: itrofimov@gmail.com

Образец ссылки на эту публикацию:

Е. П. Куршев, Е. А. Сулейманова, И. В. Трофимов. *Роль знаний в системах извлечения информации из текстов* // Программные системы: теория и приложения : электрон. научн. журн. 2012. Т. 3, № 3(12), с. 57–70.

URL: http://psta.psiras.ru/read/psta2012_3_57-70.pdf

E. P. Kurshev, E. A. Suleymanova, I. V. Trofimov. *Role of knowledge in information extraction systems.*

Abstract. The paper describes the stages of rule-based information extraction. The issues of using formalized domain knowledge at the stages of syntactic-semantic analysis and information extraction proper are touched upon. The value of enriching the extracted information with background knowledge is emphasized, and a way of implementing this function is considered. (*In Russian*).

Key Words and Phrases: information extraction, background knowledge, semantic analysis.