Подход к извлечению фактов из текста на основе онтологии[1]

An ontology-based approach to fact extraction

Сидорова Е. А. (lena@iis.nsk.su), **Кононенко И. С.** (irina_k@cn.ru)

Институт систем информатики им. А. П. Ершова СО РАН, Новосибирск

Предлагается подход к решению задачи извлечения из текста фактической информации. Используемая база знаний включает онтологию предметной области, словари предметной лексики, модель сегментации документов и схемы извлечения фактов, которые связывают термины словаря с элементами онтологии.

1. Введение

Многие информационные системы (ИнС) предъявляют весьма схожие требования к сервису анализа текста, которые сводятся к задаче преобразования слабо-структурированного текста к хорошо структурированной информации. Отличия заключаются в предметной области и структуре извлекаемых знаний.

Предлагаемая технология ориентирована на анализ документов жанра деловой прозы, для которой характерны следующие особенности: ограниченность предметной области и языка документов, наличие строгой модельной ситуации (определяемой характером автоматизации или назначением ИнС), четкость функций каждого сообщения, что позволяет сконцентрировать анализ вокруг наиболее значимых понятий предметной области. Именно такие документы являются важнейшими с точки зрения компьютерной обработки для самых различных ИнС.

Тексты деловой прозы выделяются в Национальном Корпусе Русского Языка (НКРЯ) [1] в рамках системы метапризнаков: значения признака «сфера функционирования» позволяют извлекать из корпуса и исследовать официально-деловые, публицистические, производственно-технические и учебно-научные тексты. Дополнительное привлечение признака «тип текста» позволяет дифференцировать документы по жанру (деловое письмо, научная статья и т.д.).

2. Роль онтологии предметной области при обработке текста

Основная особенность предлагаемого подхода состоит в том, что процесс анализа организован под управлением онтологии, которая расширяется за счет полученной в результате анализа информации, что, в свою очередь, является основой пополнения лингвистической базы знаний — цикличность этого процесса отмечается, например, в [2].

Отличительной чертой такого подхода является ориентация используемых лингвистических описаний на конкретные предметные знания. Словари, помимо универсальной и жанровой, содержат предметную лексику: однословные и многословные термины (словокомплексы), а также лексические шаблоны, которые представляют имена понятий и информационных объектов данной предметной онтологии (с помощью системы семантических классов, признаков и отношений). Кроме того, выбор правил анализа (как поверхностно-синтаксических, используемых для сборки словокомплексов, так и правил сборки фактов) непосредственно определяется спецификой предметной области и структурой целевой онтологической информации. Онтология определяет формат данных, которые хранятся в ИнС, и то, какую именно информацию необходимо извлекать из текста документа. Результат анализа документа представляется в виде семантической сети информационных объектов, являющихся экземплярами понятий и отношений, заданных онтологией предметной области.

[1] Работа выполняется при финансовой поддержке Президиума РАН (ИП СО РАН № 2/12 «Формальные языки и методы спецификации, анализа и синтеза информационных систем» в рамках программы фундаментальных исследований Президиума РАН № 2), РФФИ (проект № 09-07-00400).

Онтологический подход является прямым продолжением и развитием семантически ориентированного подхода к анализу и пониманию отдельного запроса и связного текста, в течение ряда лет разрабатывавшегося коллективом ИСИ СО РАН [3]. Преимущественное использование лексико-семантической информации не исключает применения частичного синтаксического анализа и синтаксических ограничений, накладываемых на семантический каркас концептуальных схем фактов (см. разделы 4.2 и 5). Известные системы различает полнота и роль синтаксического анализа в процессе извлечения фактической информации из текстов. Так, технология [4] предполагает построение полного семантико-синтаксического дерева предложения, к которому применяются шаблоны (своего рода фильтры), описывающие искомые факты. В нашем подходе, как и в [5], синтаксический анализатор применяется локально (при обнаружении ключевых единиц и их конфигураций), в частности, предусмотрено определение актантных позиций предикатных слов, ср. [6].

3. Иллюстрирующий эксперимент

Предлагаемый подход иллюстрируется на примере разработки сервиса анализа документов для информационного ресурса «Хроники СО АН».

3.1. Описание предметной области

Базовая задача анализа сообщений из архива хроник заключается в извлечении названий организаций — структурных подразделений академии наук, упоминаний персон, их ученых званий и степеней, выявление связей между персонами и организациями, а также их изменения во времени. В частности, факт переименования в онтологии отражается с помощью введения для соответствующих атрибутов множества имен с датами начала и конца их действия. Предметная онтология ИнС «Хроники СО АН» включает:

struct Data(Data_begin: data, Data_end: data);
struct Naming(Name: string, Время_действия: Data);
class *Персона* (Фамилия: Naming; Имя: string; Отчество: string; Инициалы: string; ПолноеИмя: string; Звание: domen_Звания; Степень: domen_Степени)
class *Организация* (Название: Naming; Аббревиатура: Naming)
    class Институт : Организация
    class Экспедиция : Организация **…**
relation *Сотрудник* <Персона, Организация> (Должность: domen_Должности; Дата: Data) **…**

3.2. Особенности подъязыка документов

Данный эксперимент основан на электронной коллекции документов (http://chronicle.iis.nsk.su/catalogue.aspx), единицами которой являются тексты-описания исторических событий, связанных с деятельностью Сибирского отделения Академии наук. В текстах архива излагаются наиболее существенные факты научной и научно-организационной деятельности Сибирского отделения АН СССР.

В настоящее время в архиве содержатся описания 1242 событий. Ниже приведены примеры из электронного архива, в которых выделены фрагменты, соответствующие извлекаемым объектам и фактам:

(1) ***Директору Института экспериментальной биологии и медицины СО АН СССР докт. мед. наук Е. Н. Мешалкину*** присуждена Ленинская премия за разработку новых операций на сердце и крупных кровеносных сосудах.

(2) ***В составе Сибирского отделения АН СССР*** организована самостоятельная ***Лаборатория измерительной и вычислительной электроники, которую возглавил чл.-корр. АН СССР В. Н. Авдеев***.

(3) Существование в плазме бесстолкновительных ударных волн теоретически предсказано ***чл.-корр. АН СССР Р.З. Сагдеевым (Институт ядерной физики СО АН СССР)***.

Тексты архива, являясь отрывками, извлеченными из документов различных жанров (официальных постановлений, деловых писем, газетных статей и т.п.), характеризуются небольшим размером (от одного до 4–5 предложений) и утрачивают жанровые особенности, присущие концептуальной структурной организации первоисточников. Однако тексты сохраняют лексические и синтаксические особенности деловых документов данной предметной области:

- номенклатурная лексика — наименования организаций, должностей и званий;
- шаблонная, унифицированная лексико-грамматическая структура словосочетаний, представляющих составные наименования организаций (*Институт экспериментальной биологии и медицины СО АН СССР*);
- аппозитивные конструкции, представляющие дескрипции персон, включая наименования ученых степеней, званий и составные имена (*академик Михаил Алексеевич Лаврентьев*);
- сокращения (*докт. техн. наук, доктор геол.-мин. наук, чл.-корр.);*
- однородные и скобочные конструкции, используемые для увеличения семантической емкости предложения;
- высокочастотное употребление двучленной (безагенсной) страдательной конструкции

с инвертированным порядком слов (*в состав Сибирского отделения включен Институт леса АН СССР*).

4. Представление лингвистических знаний

Лингвистическая база знаний содержит всю совокупность лингвистических знаний, необходимых для анализа текста, и включает словари предметной лексики, модель сегментации документов и схемы извлечения фактов.

4.1. Словарный компонент

При формировании словарей использовалась словарная технология, описанная в [7]. Она позволяет создавать предметные словари, которые включают однословные и многословные термины (фиксирующие частотные в анализируемом подъязыке словосочетания), а также шаблонные лексические конструкции (позволяющие определять произвольные символьные выражения, в том числе выражения, маркирующие границы сегментов). Создаваемые словари могут содержать семантические характеристики, а также накапливать статистику встречаемости терминов в текстах.

Для задачи анализа сообщений хроник были разработаны следующие словари.

1. Словарь предметной лексики, включающий имена, фамилии известных ученых, локативную лексику и термины, связанные с деятельностью научных организаций (наименования научных должностей, степеней, званий, типов мероприятий и организаций, релевантные предикаты и т.п.), — около 3 тыс. терминов.
2. Словарь лексических конструкций, включающий шаблоны дат и наименований организаций, сокращения и служебные конструкции, — около 700 шаблонов.

С целью минимизации объема ручной работы было осуществлено автоматическое начальное наполнение словарей. Для первого словаря применялись методы обучения, использующие универсальный морфологический словарь (www.aot.ru). Для второго был разработан модуль, который по набору опорных слов-классификаторов (*институт, филиал, президиум* и т.п.) и списку аббревиатур (*АН, РАН, СО РАН* и т.п.) формирует шаблоны наименований организаций вида:

[ИСИ СО РАН] =
институт**…**_систем_информатики(_
[ершова])_[СО РАН]
иси_**[СО РАН]**

Для определения основ слов из левого контекста опорного слова (когда в наименовании организации опорному слову предшествует цепочка согласованных с ним прилагательных) используется морфологический словарь. В дальнейшем эксперт вручную исправляет ошибки в аббревиатурах, устанавливает эквивалентность наименований, отмечает необязательные фрагменты, формирует иерархию шаблонов и т.п. При пополнении словаря в качестве маркеров правых границ шаблонов служат шаблоны уже известных (вышестоящих) организаций.

4.2. Модель извлечения фактов

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. Каждый факт имеет свой тип — название отношения и список его аргументов (например, отношение *Сотрудник*). Модель извлечения факта из текста должна учитывать множество языковых способов репрезентации данного отношения носителями подъязыка и обеспечивать их трансформацию в формальную структуру факта. Такую модель мы будем называть схемой извлечения факта (СИФ).

Формально, СИФ — это тройка вида < A, Res, C >, где A — множество дескрипторов аргументов факта, где дескриптором может быть тип словарной единицы, класс информационного объекта (понятие или отношение онтологии) или тип служебного факта.

Res = < t, op(t), P > — результат применения СИФ, где

1. t — задает тип элемента (класс нового объекта или один из аргументов);
2. op(t) — тип операции (создание и/или редактирование аргумента), применяемой, если выполнены ограничения C;
3. P — множество правил для формирования/редактирования объекта. Каждое правило ставит в соответствие атрибуту результирующего объекта либо точное значение, либо значение атрибута одного из аргументов.
4. C — множество ограничений, накладываемых на характеристики аргументов факта. Выделяются следующие ограничения:
5. условия на морфологические и семантические характеристики аргументов схемы (например, arg1.Падеж = рд, arg1.SemClass=Лок)
6. ограничение синтаксической сочетаемости вершин синтаксических групп, реализующих аргументы схемы (например, Synt = Согл(число, падеж), см. схему 2),
7. структурно-текстовые ограничения на взаиморасположение аргументов в тексте: позиция

аргументов относительно друг друга, тип контактности (например, Contact=Common — аргументы могут разделяться в тексте знаками препинания и/или незначимыми словами), тип сегмента.

Приведем пример типичной схемы:

Scheme Персона_с_инициалами **(1)**
arg1: Term::*ФИО*(фио-тип: *фам*)
arg2: Term_lex::*инициалы*()

Condition Position = preposition_priority, Contact = absolute

⇨ Object::*Персона*(Фамилия: arg1.Name, Инициалы: arg2.Value)

Данная схема имеет два аргумента, которые описывают элементы из словарей разного типа (словарь предметной лексики и словарь шаблонов). Позиционное ограничение определяет контактность терминов в тексте, а также указывает на тот факт, что расположение фамилии справа от инициалов является приоритетным в случае, когда есть альтернатива (это позволяет корректно обрабатывать ситуации вида *Г. Петров И.И.*). В результате будут формироваться объекты класса Персона с двумя определенными атрибутами (*Фамилия* и *Инициалы*). Эта схема применима к любому контексту типа *ФИО*, в котором фамилия сопровождается инициалами.

5. Извлечение информации из текста

Процесс обработки текста включает следующие этапы: графематический анализ, лексический анализ, сегментация, морфологический анализ, сборка фактов и формирование контента документа.

5.1. Сегментация

Предусмотрена возможность осуществлять два вида сегментации текста — первичную (логическую) и жанровую [8]. В процессе первичной сегментации производится разбиение линейного текста на строковые объекты, оформленные как сегменты и упорядоченные в соответствии с их встречаемостью в тексте. В рассматриваемом примере документы не имеют выраженной жанровой структуры, поэтому процесс сегментации порождает только логические сегменты: абзац, предложение, клауза и т.п.

Разбиение на сегменты используется при сборке фактов, где, при наличии соответствующего структурного ограничения, на вход алгоритму подается не весь текст целиком, а только фрагмент текста. В этом случае алгоритм сборки фактов запускается столько раз, сколько найдено требуемых сегментов.

5.2. Сборка фактов

Процесс извлечения фактов из текста хроник базируется на схемах извлечения фактов, при формировании которых максимально полно учитываются различные способы выражения в текстах объектов и отношений предметной онтологии.

5.2.1. Извлечение объектов.

Извлечение из текста объекта класса Персона, представленного именной группой типа ФИО, демонстрируется схемой 1 в разд. 4.2.

При вводе в текст объекта класса Персона могут быть указаны ученые степень и звание. В этом случае дескрипция объекта, как правило, представляет собой аппозитивную именную группу, в которой к группе типа ФИО примыкают, чаще в препозиции, фрагменты (сокращения или согласованные с ФИО по числу и падежу именные группы), реализующие Степень и Звание. Схема 2 демонстрирует определение значения атрибута Звание, характеризующего объект класса Персона. С помощью аналогичной схемы извлекается значение атрибута Степень.

Scheme Персона_звание : segment *Предложение* **(2)**
arg1: Object::*Персона*()
arg2: Term::*Звание*()

Condition Position = postposition_priority, Contact = common, Synt = Согл (число,падеж)

⇨ arg1(Звание: arg2.Name)

В текстах хроник отмечаются ситуации нереферентного употребления имен собственных, когда идентифицированный в тексте фрагмент ФИО не вводит конкретного объекта класса Персона: *А. П. Виноградова* в контексте *Институт геохимии* ***им. А. П. Виноградова***; упоминание персон в позиции актанта (С рд) предиката *в честь, памяти,* а также *имя* (в контексте *присвоить, получить, носить*), как *акад. А. П. Виноградова* в примере (4). Это случаи, в которых может иметь место тот или иной вариант локальной неоднозначности (наименование персоны vs. фрагмент наименования организации).

(4) Институту геохимии СО АН СССР присвоено имя выдающегося советского ученого акад. А. П. Виноградова.

В первом случае омонимия снимается уже на уровне сборки лексических шаблонов объектов: подстрока *А. П. Виноградова* входит в лексическую конструкцию, реализующую шаблон наименова-

ния объекта класса Организация. Таким образом, шаблон, охватывающий подстроку, соответствующую группе ФИО, имеет отрицательную видимость и объекта не создает.

В остальных случаях снятие неоднозначности требует не только лексического анализа, но и обработки на этапе сборки фактов. Идентификация объекта Персона в указанной актантной позиции позволяет изменить статус найденного объекта на нереферентный (схема 3), означающий, что в БД ему не сопоставляется никакой конкретный объект, а если такой объект уже присутствует в БД, то ему не добавляется никакой новой информации. Одновременно инициируется формирование служебного факта Именование: отношение Именование не представлено в онтологии хроник, но данный факт позволяет извлечь связанную с (пере)именованием дату.

Scheme Имя_Персоны: segment *Клауза* **(3)**
arg1: Term::*Предикат_Имя*()
arg2: Object::*Персона*(Падеж: рд)

Condition Position = preposition, Synt = Упр(arg1, рд)

⇨ arg2(Visibility: false), Fact::*Именование*(second: arg2)

5.2.2. Извлечение отношений

Рассмотрим пример извлечения из текста отношения Сотрудник, аргументами которого являются объекты классов Персона и Организация. Это отношение в текстах хроник представляется как факт сотрудничества персоны в организации в некоторой должности, которая может быть определенной (*директор, научный сотрудник*), неопределенной (*сотрудник, ученый*) либо недоопределенной должностной ролью 'первого лица организации' (*глава, руководитель*).

Различные способы репрезентации отношения Сотрудник сводятся к двум основным вариантам (ср. [5]).

1. В первом варианте для репрезентации отношения Сотрудник используется непредикативная конструкция — связь объектов Организация и Персона реализуется в синтаксических рамках именной группы.

В примере (1) реализована наиболее типичная схема: именная группа <Должность + Организация> (построенная по схеме С+Срд) и примыкающая согласованная (по числу и падежу) именная группа, реализующая объект Персона. Вся конструкция неразрывна, хотя возможны различные варианты взаимного расположения групп и сегментации: все компоненты представлены аппозитивной конструкцией в одном сегменте, как в примере (1); все или некоторые компоненты принадлежат разным сегментам, в том числе возможны сегменты скобочного типа, как в примере (3). Заметим, что скобочная структура не предполагает согласованности группы в скобках с остальными компонентами.

В другой типичной схеме связь групп Организация и Персона реализуется через предложно-падежное примыкание (*во главе с, под руководством*).

2. Во втором варианте связь объектов Организация и Персона реализуется предикативно, с помощью эксплицитных глагольных предикатов, включая лексемы, непосредственно репрезентирующие отношение сотрудничества (*принять, уволить, утвердить, назначить, избрать, работать*), предикаты, вводящие должностную роль 'первого лица' (*возглавлять, руководить*), а также связочные глаголы (*быть, становиться, являться, занимать, находиться*). В реализации этой связи в актантной позиции регулярно используется группа Должность (*принять в институт техником, назначен на пост директора института*).

В качестве подкласса предикатов, репрезентирующих отношение Сотрудник, в словаре имеются глаголы *руководить, возглавлять*, в семантике которых синкретично выражено значение должностной роли первого лица организации. В процессе анализа примера (2) данная информация извлекается посредством схемы 4, которая применима без каких-либо ограничений на морфологический класс (часть речи) предиката.

Scheme Предикат_ Сотрудник_ первое_лицо **(4)**
arg1: Term::*СотрудникПЛ*()

⇨ Relation::*Сотрудник*(status1:«missing», status2:«missing», Должность_роль: «первое лицо»)

Далее используются схемы, применимые к произвольному предикату класса Сотрудник, представленному в любой глагольной форме, возможной в позиции вершины клаузы (личный глагол, причастие, инфинитив и т.п.). Ограничение синтаксической сочетаемости проверяет согласованность грамматических признаков вершин синтаксических групп, реализующих аргументы схемы, в соответствии со стандартными правилами согласования (Согл) и управления (Упр i). Конкретный вид ограничения определяется значениями морфологических характеристик аргументов. В примере (2) отношение (arg1 схемы) представлено личным глаголом (*возглавил*), а 1-му семантическому актанту отношения (Персона) соответствует подлежащее. Это означает применение ограничения (arg2.Падеж=им и Согл (Род, Число)). В ситуации со страдательным причастием (*возглавляемый*) применяется ограничение (arg1.Упр2=arg2.Падеж), в данном случае это arg2. Падеж=тв. Соответственно, действительное причастие в позиции arg1 схемы означало бы ограничение вида Согл (Род, Число, Падеж).

При отсутствии одного из актантов в пределах клаузы (в примере (2) это Организация) наличие в ней местоименного заместителя (*который*) означает применимость схем разрешения анафоры. Восстановление антецедента происходит в два этапа. Сначала путем проверки падежной формы местоимения уточняется факт анафорической замены второго семантического актанта (arg2.Упр2 = arg1.Падеж). Затем происходит собственно установление антецедента (схема 5, требующая согласования местоименного заместителя *который* с объектом, претендующим на роль антецедента), что и завершает процесс извлечения факта на основе отношения Сотрудник из текста (2).

Scheme Сотрудник_Антецедент_Орг: segment *Предложение* **(5)**
arg1: Relation::*Сотрудник*(status2: «antecedent — left segment»)
arg2: Object::*Организация*()

Condition Position = postposition, Synt = Согл(Род, Число)

⇨ arg1(second: arg2, status2 = «complete»)

5.3. Генерация информационных объектов

После извлечения фактов из текста осуществляется генерация информационных объектов, соответствующих найденным фактам. На данном этапе происходит взаимодействие с БД системы с целью уточнения объектов и их характеристик:

- Слияние референтных объектов на основе результата процедуры поиска соответствующих объектов в БД (если двум объектам сопоставился один объект БД, то делается вывод о тождестве данных объектов);
- Уточнение неявно выраженных характеристик, например, если в отношении Сотрудник атрибут Должность_роль = *«первое лицо»*, то определяется значение атрибута Должность на основании информации о типе Организации и названии руководящей должности для данного типа:

Сотрудник.Должность_роль = «*первое лицо*» +
Организация.Тип = «*лаборатория*»
⇒ Сотрудник.Должность = «*заведующий лабораторией*».

Для того чтобы сформировать результирующий контент документа, необходимо:

1. обеспечить контроль корректности значений атрибутов информационных объектов, полученных в результате анализа;
2. идентифицировать полученные объекты, т.е. заполнить ключевые атрибуты и сопоставить с объектами базы данных ИнС (если такие существуют);
3. добавить объекты в информационное пространство системы и связать их с документом.

При редактировании объекта в БД могут возникать противоречия между старыми и новыми значениями его характеристик. Для решения данной проблемы была выбрана стратегия сохранения всех данных с указанием даты.

6. Заключение

Проводимое исследование по извлечению фактической информации из текстов хроник ограничено отношениями между персонами и организациями, а также между организациями, с учетом локативных и временных характеристик этих отношений. Схемы описывают реализации фактов в рамках именных и предикативных конструкций, которые могут осложняться анафорой и однородными группами (*члены-корреспонденты АН СССР А. А. Трофимук, М. М. Шемякин*).

В ближайшее время планируется расширение лингвистической базы знаний средствами репрезентации объектов и отношений онтологии и совершенствование собственно лингвистического анализа в аспекте нерешенных проблем. В частности, одним из источников ошибок являются составные наименования объектов и атрибутов при установлении референции (*Председателем РИСО утвержден акад. С. Л. Соболев, его заместителем — акад. А. Л. Яншин*. — антецедент не находится) и обработке сочинительного сокращения (*с институтами: геологии, горного дела, биологическим, радиофизики и электроники* — 4 или 5 организаций?). Требуют внимания и ситуации со снятым или нереальным отношением (*бывший директор, не переизбран на пост ректора,* предполагается принять на должность).

В дальнейшем предполагается проведение эксперимента на широкой тестовой базе (новости и постановления Президиума СО РАН).

Литература

1. *Национальный* Корпус Русского Языка www.ruscorpora.ru.
2. *Nedellec C. and Nazarenko C.* Ontology and Information Extraction: A Necessary Symbiosis // Ontology Learning from Text: Methods, Evaluation and Applications. Buitelaar P., Cimiano P. and Magnini B.(eds .), IOS Press Publication:2005.
3. *Нариньяни А. С.* Автоматическое понимание текста — новая перспектива // Труды международного семинара Диалог'97 по компьютерной лингвистике и ее приложениям. М.: 1997. С. 203–208.
4. *Ермаков А.Е.* Автоматическое извлечение фактов из текстов досье: опыт установления анафорических связей // Труды международной конференции Диалог' 2007 «Компьютерная лингвистика и интеллектуальные технологии». М.: Наука, 2007.
5. *Гершензон Л. М., Ножов И. М., Панкратов Д. В.* Система извлечения и поиска структурированной информации из больших текстовых массивов СМИ. Архитектурные и лингвистические особенности // Труды международной конференции Диалог'2005 «Компьютерная лингвистика и интеллектуальные технологии». М.: Наука, 2005. С. 97–101.
6. *Азарова И. В., Гребеньков А. С., Ландо Т. М.* Использование маркеров актантных позиций при анализе деловых текстов для расширения логической схемы предметной области // Труды международной конференции Диалог' 2008 «Компьютерная лингвистика и интеллектуальные технологии». М.: РГГУ, 2008. Вып. 7 (14). С. 11–16.
7. *Сидорова Е. А.* Многоцелевая словарная подсистема извлечения предметной лексики // Труды международной конференции Диалог' 2008 «Компьютерная лингвистика и интеллектуальные технологии». М.: РГГУ, 2008. Вып. 7 (14). С. 475–481.
8. *Кононенко И. С., Сидорова Е. А.* Обработка делового письма в системе документооборота // Труды международного семинара Диалог'2002 по компьютерной лингвистике и ее приложениям. Протвино: 2002. Т. 2. С.299–310.