ЛИНГВИСТИЧЕСКИЕ И АЛГОРИТМИЧЕСКИЕ АСПЕКТЫ ВЫДЕЛЕНИЯ ОБЪЕКТОВ И СВЯЗЕЙ ИЗ ПРЕДМЕТНО-ОРИЕНТИРОВАННЫХ ТЕКСТОВ

LINGUISTIC AND ALGORITHMIC ASPECTS OF OBJECT EXTRACTION FROM SUBJECT TEXTS OF NATURAL LANGUAGE

Кузнецов И.П. *(igor-kuz@mtu-net.ru),* ***Мацкевич А.Г.***
Институт проблем информатики РАН

Рассматриваются проблемы построения одного класса семантико-ориентированных лингвистических процессоров, выделяющих из текстов естественного языка информационные объекты и их связи. Настройка процессоров на предметную область осуществляется за счет лингвистических знаний. Анализируются опыт использования таких процессоров для формализации текстов в различных предметных областях: криминалистики (сводки происшествий, обвинительные заключения и др.), СМИ (документы о террористической деятельности), кадры (автобиографии, резюме). Рассматриваются особенности каждой такой области: наборы выделяемых объектов, необходимость их идентификации, связи, а также имеющие место сокращения, разделительные знаки, специфика языковых конструкций и др. Такие особенности были учтены при разработке лингвистических знаний.

Введение

Лавинообразный рост потока документов, получаемых пользователями через различные информационные каналы (в том числе из сети Интернет), требует новых решений. Большая часть таких документов имеет вид текстов естественного языка (ЕЯ). Во многих случаях человек не в силах прочитать и осмыслить даже малую часть того, что ему предлагается. Существующие средства могут оказать помощь, но для этого требуется предварительная работа по формализации текстов или запросов.–

В тоже время большинство конкретных пользователей – это люди, которые интересуются конкретными вещами. Например, следователю важны фигуранты, их место жительства, телефоны, криминальные события, даты и др. Специалиста по кадрам интересуют организации, где человек работал, кем и когда это было. Другие люди вылавливают из СМИ информацию о странах, влиятельных лицах, катастрофах и др. Будем называть интересующую пользователя конкретную информацию – *информационными объектами*.

Отсюда следует необходимость построения нового класса информационных систем, которые должны учитывать интересы конечного пользователя и быть ориентированы на извлечение из текстов [1-3] информационных объектов. Эта проблема в настоящее время находится в фокусе внимания отечественных и зарубежных исследователей и разработчиков [4-19].

В данной статье рассматривается класс таких систем, основанных на использовании специальных лингвистических процессоров (ЛП) и технологии *баз знаний* (БЗ). Лингвистические процессоры необходимы для глубинной обработки текстов с выявлением информационных объектов и связей. На основе последних формируются структуры знаний, которые образуют БЗ. Будем называть такие процессоры семантико-ориентированными. Их особенность – в наличии *лингвистических знаний* (ЛЗ), организованных таким образом, чтобы учитывать лексические и семантические особенности ЕЯ при формировании структур знаний [1,14].

На уровне БЗ становится возможным более полно учитывать потребности пользователя в следующих направлениях. Во-первых, за счет использования обратных лингвистических процессоров для формирования отчетов, заполнения требуемых форм или таблиц, в том числе, реляционных БД. Во-вторых, за счет поддержки экспертной компоненты, обеспечивающей пополнение данных аналитическими результатами, полученными при обработке структур знаний. И, в третьих, за счет организации различных видов поиска: поиска конкретных объектов, поиска похожих объектов, поиска по связям и др. Такие виды поиска относятся к "семантическим", так как осуществляется не на уровне слов или словоформ, а на уровне структур знаний из БЗ. Будем называть системы подобного типа семантико-ориентированными.

На протяжении последних 15 лет на базе исследований, проводимых в ИПИ РАН, были разработаны семантико-ориентированные системы и ЛП для формализации текстов ЕЯ и их аналитической обработки в различных предметных областях: криминалистики (сводки происшествий, обвинительные заключения и др.), СМИ (документы о террористической деятельности), кадры (автобиографии на русском и английском языках). Это комплексные системы ДИЕС, ИКС, "Аналитик, "Криминал", Lingvo-Master [12-17]. В данной работе будет идти речь об особенностях этих систем (используемых в них ЛП и ЛЗ), которые определяются задачами и спецификой текстов ЕЯ.

1. Область криминалистики.

Потоки документов в криминальной милиции – это сводки происшествий, справки по уголовным делам, обвинительные заключения и др. В этих документах содержится много конкретной информации, касающейся фигурантов, их деяний, орудий преступления и др. Основные задачи – различные виды поиска. Отметим, что объемы ежемесячной новой информации подобного типа исчисляются десятками и сотнями мегабайт. Никто не может все это прочитать и держать в голове.

Полнотекстовые базы данных не решают проблемы, так как при работе с текстами на ЕЯ дают много шумов (лишних документов) и потерь. Причина этому – особенности русского языка: наличие словоформ, свободный порядок слов. Одно и тоже можно выразить множеством различных способов. Слова запроса могут быть разбросаны по тексту документа и относиться к различным сущностям. Для устранения этих недостатков вводят критерии близости слов, обрезают окончания словоформ, проводят индексирование нормализованных слов и др. Но и это кардинально не решает проблемы.

Другой вариант – это использование реляционных БД. Но для этого требуются трудоемкая работа специально обученных людей по формализации текстов на ЕЯ: выделение из текстового документа (происшествия) лиц, адресов, дат,... и заполнение соответствующих таблиц БД. При больших потоках документов это сделать крайне трудно.

В связи с этим в конце 90-х годов была разработана система "Криминал" [12,13]. Ее особенность – автоматический анализ текстов с выделением необходимого набора информационных объектов. Система "Криминал" отлаживалась на 500 тыс. происшествий из сводок ГУВД г. Москва и по основным объектам удалось добиться хороших результатов: коэффициент шумов в компонентах (лишних слов в объектах) – не более 1-2% и потерь (отсутствие нужных слов) – не более 1%.

Основные выделяемые объекты (потери должны быть минимальными):

- лица (по ФИО) с их особенностями (преступник, потерпевший);
- словесное описание лиц, их приметы;
- адреса, почтовые атрибуты;
- даты;
- оружие с атрибутами;
- номера телефонов, факсов, е-майлов с их стандартизацией;
- средства транспорта с выделением марки машины, государственного номера, цвета и других атрибутов;
- паспортные данные и другие документы с их атрибутами;
- взрывчатые вещества и наркотические вещества;
- отделения милиции;
- сотрудники милиции.

Второстепенные объекты (потери допустимы):

- организации;
- должности;
- количественные характеристики (сколько лиц или других объектов принимали участие в том или ином событии);
- номера счетов, суммы денег с указанием типа валюты;

Связи:

- события (криминальные, террористические, поломки изделий и др.) с указанием участия в них информационных объектов;
- время и место событий;
- связи между различными типами информационных объектов (кем работает лицо в той или иной организации, по какому адресу проживает, в каких событиях принимал участие совместно с другими объектами и т.д.).

Некоторые из трудностей извлечения объектов из текстов заключаются в следующем. Во-первых, трудности, связанные с особенностями русского языка. Это свободный порядок слов, наличие омонимии и полисемии,

разнообразие языковых форм для выражения одного и того же. Например, какое-либо событие можно выразить с помощью глагольных форм, отглагольных существительных, причастных оборотов и др. Их нужно приводить к одному виду.

Во-вторых, наличие (особенно в сводках происшествий) большого количества сокращений, которые нужно расшифровывать путем анализа контекста. Например, Г. может означать ГОД, ГОРОД, ГОС. и др.

В-третьих, много умолчаний. Например, после фигуранта пишется его адрес, год рождения и другие данные. Их нужно связывать с фигурантом.

Еще одна задача – это идентификация объектов (фигурантов) по всему тексту, использование для этих целей указательных местоимений, кратких имен, анафорических ссылок. Это особенно необходимо для обвинительных заключений, где одно и тоже лицо упоминается многократно (различными способами именования) по всему документу.

С учетом трудностей и в соответствии с задачами был разработан лингвистический процессор системы "Криминал", осуществляющий нормализацию слов, их группировку с формированием объектов, идентификацию объектов и установление связей. В результате по каждому документу ЕЯ автоматически строились семантическая сеть, называемая содержательным портретом документа. Последние – это структуры знаний, которые составляют БЗ и на основе которых были реализованы различные виды семантического поиска: поиск по признакам и связям, поиск связанных объектов на различных уровнях, поиск похожих фигурантов и происшествий, поиск по приметам (с использованием онтологий).

Поддерживается **экспертная компонента** – для классификации происшествий по каталогам криминальной милиции: "Вид преступления", "Способ совершения преступления" и др. Результат вводится в содержательный портрет. Имеется полный набор настроек на предметную область.

2. Задачи кадровых агентств

Одна из важных проблем кадровых и рекрутинговых агентств связана автоматической обработкой автобиографических данных, заявок на работу (резюме), написанных в достаточно произвольной форме – в виде текстов ЕЯ. Такие тексты содержат сведения о человеке: ФИО, год рождения, адрес, время и место учебы с указанием наименования учебного заведения и др. Требуется их автоматическая формализация с выделением информационных объектов и их отображением на поля заданной анкеты или сайта. Тогда становится возможным использование типовых средств баз данных для решения пользовательских задач. Во многих агентствах такая формализация делается вручную: специально подготовленными людьми, или же самим человеком, которому предлагается ввести его сведения в указанные поля по требуемой форме. Это достаточно трудоемкая работа.

В качестве основы для автоматизации этих работ был взят лингвистический процессор системы "Криминал". Однако, он был доработан в соответствии с особенностями предметной области [17]. Во-первых, это необходимость выделения другого набора объектов и связей. Во-вторых, их деление на группы. Например, деление объектов (организаций, дат и др.) на те, которые относятся к учебе или к профессиональной деятельности или к курсам. В-третьих, необходимость использования экспертных систем для пополнения данных, которые заданы в неявном виде. Будем называть такие данные экспертными объектами.

Основные объекты:

• лицо, составляющее заявку (как правило, в самом начале заявки);
• дата рождения или возраст;
• E-mail;
• почтовый адрес;
• домашний телефон;
• мобильный телефон;
• рабочий телефон;
• личная интернет-страница;
• желаемая должность;

УЧЕБА

• название учебного заведения;
• факультет (специальность);
• диплом (степень);
• начало учебы (дата);
• окончание учебы (дата);

ПРОФЕССИОНАЛЬНЫЙ ОПЫТ
• начало работы (дата);
• окончание работы (дата);
• название организации;
• занимаемая должность;
• обязанность, функции, достижения;

КУРСЫ (обучение)
• проводящая организация;
• название курсов;
• диплом (сертификат);
• начало курсов;
• окончание курсов.

Экспертные объекты:
• пол;
• образование (среднее, высшее и др.);
• профессиональная область (по заданной классификации);
• специализация (по заданной классификации);
• опыт работы (суммируется количество лет);
• регион (вычисляется по адресу);
• знание языка (по степени владения).

Выделение многих из этих объектов потребовало лишь доработки лингвистических знаний (ЛЗ). Однако, особенности текстов и решаемые задачи потребовали усиления возможностей самого ЛП. Это было вызвано следующими факторами. Во-первых, разнообразием форм ЕЯ, с помощью которых выражаются даты и временные интервалы. Например, даты могут быть в сокращенной форме (авг.05), в виде дробных чисел (09.99 г.), разного рода специальных знаков или кавычек (09/99 или 09’1999) и т.д. Интервалы: 15.05-01.12.99 или май-июнь 06 и др. Трудности вызывали их путаница с дробными числами, отсутствие ключевых слов типа г. (год) и др. Более того, одним из требований было приведение дат к стандартному виду – расшифровка сокращений.

Во-вторых, определенные трудности вызывали задачи деления объектов на типы и правила их компоновки. Например, сравнительно часто в резюме на ЕЯ такие объекты как организации (где человек работал или учился), должности, периоды работы и основные обязанности ставятся в достаточно произвольной последовательности. Если период работы в какой-либо организации записан в конце и далее идет другая организация, то нужно уметь определять, куда отнести этот период. Периоды, даты или другие организации (например, заказчики проекта) могут стоять и внутри текста описания работы, что вызывает дополнительные трудности. Человеку по смыслу проще понять, что к чему относится. Значительно труднее выработать формальные критерии разделения и соотнесения, которые бы давали допустимое количество шумов и потерь. В связи с этим в ЛП были введены специальные средства, которые, опираясь на даты (или организации), осуществляли поиск связанных с ними объектов.

В-третьих, многие пользователи создавали свои резюме на основе документов, взятых из различных таблиц, форм. Как следствие, отсутствие знаков препинания (точек), наличие спецзнаков, остающихся после перекодировки текстов. Все резюме (если не было пробельных строк) воспринималось как одно предложение.

В связи с этим в блок морфо-лексического анализа были введены специальные средства настройки – правила для выделения предложений [19]. Например, если слово-глагол написано с большой буквы и стоит вначале строки, то это начало предложения. Таких правил множество, в том числе такие, которые учитывают роль спецзнаков, разделительных символов и др.

В-четвертых, для получения экспертных данных (объектов) в ЛП были встроены экспертные системы (ЭС), которые, например, на основе анализа содержательных портретов соотносят документ к определенной категории (пункту классификатора), или же на основе имеющегося описания определяют степень владения иностранным языком и т.д. Если такая информация указана в исходном тексте в явном виде, то экспертной оценки не требуется.

В системе реализовано два типа оболочек для ЭС. Первая основана на весовых коэффициентах слов, соответствующих определенной категории. Вторая – на наличии слов в информационных объектах.

В ЭС первого типа с каждой категорией связываются слова с указанием их весов. Такие веса являются результатом статистического анализа эталонных документов (проанализированных человеком), т.е. предполагается этап обучения.

В ЭС второго типа с каждой категорией связываются характеризующие слова или пары слов (словосоче-

тания), которые берутся из фрагментов, соответствующих информационным объектам указанного типа. Одно и тоже слово или словосочетание может соотноситься лишь с одной категорией.

И наконец, последнее – это необходимость в обратном ЛП. Обратный ЛП служит для преобразования объектов в компоненты ЕЯ и для их отображения на поля анкеты или сайта. Этот процессор имеет свои лингвистические знания, с помощью которых задается последовательность выдачи рубрик (полей) и какими объектами они должны заполняться. Для выделения таких объектов служат их имена (ОРГ_, РАБ_,...), а также связи, заданные в содержательном портрете. Для каждого выделенного объекта строится его описание – из входящих в него нормализованных слов. Далее, по объекту находится его предложение. За счет средств позиционирования находится место предложения в тексте, т.е. интервал от байта до байта. По описанию объекта в этом интервале ищется кусок предложения, соответствующий объекту. Этот кусок и выдается в качестве результата.

3. Документы СМИ о террористической деятельности

Проблема информационной поддержки борьбы с терроризмом в современном мире стоит очень остро и привлекает внимание исследователей, однако работающие системы извлечения знаний для этой области только начинают создаваться [18].

Основная задача – выделение из потока сообщений СМИ тех документов, которые относятся к террористической деятельности, с последующим анализом этих документов. В качестве основы нами был взят лингвистический процессор (ЛП) системы “Криминал”. Он был доработан в соответствии с особенностями предметной области и задач. В ЛП были дополнительно введены следующие информационные объекты:
террористические группы и организации (Terrorizm);
участник террористические группы с указанием его роли (лидер, главарь и др.);
вооруженные силы, используемые для борьбы с терроризмом (Military_Force);
интервалы времени (см. п. 2).

Были разработаны лингвистические знания (ЛЗ) для выделения этих объектов. В соответствии со спецификой текстов ЛЗ были дополнены новыми правилами выделения объектов, например, выделение места события по формам “в 25 км. От Кабула” или “лагерь близ города Умма” и т.д. Особые трудности вызывало выделение арабских составных имен с их элементами абд (раб), Абу (отец), Ибн или Бен (сын) и др. Они не укладываются в формат европейских стандартов. Например, Абд ар-Расул, бен-Ахмад. Соответственно, усложняется ФИО. Для известных террористов, как правило, используются сокращенные имена, например, Бен Ладен (вместо Усама Бен Ладен), Басаев (Шамиль Басаев) и др. В ЛП были введены специальные средства их идентификации.

Как и в предыдущих случаях, при выделении объектов учитываются возможные варианты называния объекта в тексте, в том числе, краткой форме. Типовые объекты (ФИО, даты, адреса, виды оружия и др.) приводятся к одному (стандартному) виду. Осуществляется идентификация объектов с учетом кратких наименований (например, отдельных фамилий или имен с ФИО), анафорических ссылок (указательных и личных местоимений, например, “Этот человек”, “Он ...”), определений (например, “Мэр Москвы Лужков” идентифицируется с последующими словами “мэр”, “Лужков”). Для выделения событий и связей проводится анализ глагольных форм, а также причастных и деепричастных оборотов.

В тоже время основная задача использования ЛП отличалась от предыдущих случаев – это необходимость работы (в качестве отдельного модуля) в рамках комплексных систем сбора и обработки информации. Обмен – через XML-файлы [20]. В связи с этим был разработан обратный ЛП, который на основе содержательных портретов строит XML-файлы (см. Приложение 1).

Таким образом, на входе процессора (ЛП) – текст ЕЯ, а на выходе – XML-файл, где представлены все выделенные объекты и связи с указанием источников. Такой ЛП под названием Sematix поставляется в виде SDK-модуля. Работает в среде WINDOWS. Может быть перекомпилирован для работы под LINUX.

Процессор Semantix является отторгаемым модулем и может быть использован вне упомянутых систем для типовых задач аналитических служб. Имеются средства настройки на объекты других типов – за счет лингвистических знаний или словарей.

Пример работы Semantix приведен в Приложении 1. Дадим некоторые пояснения. Каждый объект имеет следующую структуру:

```
<OBJECT ID=”7” TYPE=”Organization”>
  <ARG CONST=”ШТАБ“ />
  <ARG CONST=”КВАРТИРА” />
  ...
  <SOURCE> Штаб квартиру оппозиционной группы</SOURCE>
</OBJECT>
```

где ID=”10” – идентификационный номер объекта, а TYPE=”Organization” – его тип. Также дается компонента текста, соответствующая объекту. Отношения объектов и их участие в действиях представлены через ссылки REF=... Например, с помощью конструкции:

```
<ACTION ID=”15” TYPE=”УДАР“>
  <ARG CONST=”НА“ />
  <ARG REF=”7” />
</ACTION>
```

где представлено “один из ударов пришелся на штаб-квартиру оппозиционной группы”. Для каждого объекта или действия дается ссылка на предложение.

В процессоре Semantix использована достаточно универсальная конструкция XML-файла: один объект (через ссылку) может включать в себя другой объект. Свойства даются как аргументы. В случае необходимости указывается тип атрибута. Например, <ATTR TYPE=”YEAR” VALUE=”2003” /> – указан год и т.д. XML-файл имеет полный набор информации, необходимой для использования в различных комплексных системах.

Заключение

Объектно-ориентированные лингвистические процессоры могут быть использованы в различных областях приложений, где требуется извлечение полезной информации из текстов естественного языка. При этом, процессоры, которые описаны в данной работе, обладают рядом существенных преимуществ. Недавно появившиеся системы Интегро Онтос, Арион и др. (насколько известно авторам) устойчиво выделяют лишь объекты нескольких типов. Как правило, это лица, организации, даты, адреса (к сожалению, в итак немногочисленных публикациях в этой области много рекламы, общих слов и мало реальных данных).

В процессорах Semantix, Lingvo-Master и системы “Криминал”, выделяется до 40 типов объектов с высокой точностью и минимальными шумами. Например, система “Криминал” отлаживалась на 500 тыс. происшествий из сводок ГУВД г. Москва и показала уникальные результаты: коэффициент шумов по основным объектам (объект неправильно найден) – не более 1% и потерь (объект имеется в тексте, но не найден) – менее 0.5%. Процессор Semantix отлаживалась на меньшем количестве документов, где речь шла о террористической деятельности, и поэтому в нем может быть больше шумов и потерь. Но это быстро устранимо.

Дело в том, что учесть все, что может встретиться в текстах ЕЯ, не представляется возможным. Поэтому чрезвычайно важны, во-первых, представительный набор тестовых документов, и во-вторых, средства отладки. или настройки лингвистического процессоры: наличие трассировок различного уровня, средств быстрой корректировки и подстройки лингвистических знаний. В наших системах имеется весь комплекс таких средств, которые обеспечивают быструю настройку на приложения (в том числе, ввод новых объектов и связей) с учетом требований заказчика [19].

Отметим, что в упомянутых процессорах объекты приводятся к стандартному виду (например, ФИО, адреса, даты) с указанием типов компонент. Проводится достаточно глубокий анализ предложений с выявлением глагольных форм, а также с идентификацией объектов по всему тексту. Обеспечивается анализ сложных языковых конструкций: форм с отглагольными существительными, причастными, деепричастными оборотами, однородными членами и др. Поддерживается экспертная компонента. Процессор Semantix может быть использован как отторгаемый (независимый) модуль.

В настоящее время разработан англоязычный вариант объектно-ориентированного лингвистического процессора Semantix [15,16,19].

Список литературы

1. Кузнецов И.П. Семантические представления // М. Наука. 1986г. 290 с.
2. Cunningham, H. Automatic Information Extraction // Encyclopedia of Language and Linguistics, 2cnd ed. Elsevier, 2005.
3. Han J. and Kamber, M. Data Mining: Concepts and Techniques // Morgan Kaufmann, 2006.
4. FASTUS:a Cascaded Finite-State Trasducerfor Extracting Information from Natural-Language Text. // AIC, SRI International. Menlo Park. California, 1996.
5. Ferrucci, D. and Lally, A. UIMA: an architectural approach to unstructured information processing in the corporate research environment // Natural Language Engineering 10 (3/4), 2004, 327–348.
6. Byrd, R. and Ravin, Y. Identifying and Extracting Relations in Text // 4th International Conference on Applications of Natural Language to Information Systems (NLDB). Klagenfurt, Austria, 1999.

7. Popov, B. et al. KIM – A Semantic Platform for Information Extraction and Retrieval // Journal of Natural Language Engineering, 10(3-4), 2004, pp. 375-392.
8. Doddington, G. et al. Automatic Content Extraction (ACE) program – task definitions and performance measures // Fourth International Conference on Language Resources and Evaluation (LREC), 2004.
9. Han, J., Pei Y. Yin, and Mao, R. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach," // Data Mining and Knowledge Discovery, 8(1), 2004, pp. 53–87.
10. Dong, G. and J. Li. Efficient mining of emerging patterns: Discovering trends and differences // Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and DataMining, S. Chaudhui and D. Madigan, editors, ACM Press, San Diego, CA, 1999, pp. 43–52.
11. Добров Б.В., Лукашевич Н.В. Онтологии для автоматической обработки текстов: Описание понятий и лексических значений // Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. конференции Диалог'06, Бекасово, 31 мая – 4 июня 2006 г., 2006, стр. 138-142.
12. Кузнецов И.П. Методы обработки сводок с выделением особенностей фигурантов и происшествий // Труды международного семинара Диалог-1999 по компьютерной лингвистике и ее приложениям. Том 2. Таруса 1999.
13. Кузнецов И.П., Мацкевич А.Г. Система извлечения семантической информации из текстов естественного языка // Труды межд. Семинара Диалог 2001 по комп. лингвистике и её приложениям: Т.2. Москва, Наука 2002.
14. Кузнецов И.П., Особенности обработки текстов естественного языка на основе технологии баз знаний // Сб. ИПИ РАН, Вып.13, 2003 г. стр. 241-250.
15. Kuznetsov, I., Kozerenko, E. The system for extracting semantic information from natural language texts // Proceeding of International Conference on Machine Learning. MLMTA-03, Las Vegas US, 23-26 June 2003, p. 75-80.
16. Кузнецов И.П., Мацкевич А.Г. Англоязычная версия системы автоматического выявления значимой информации из текстов естественного языка // Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям "Диалог 2005", Звенигород, 2005.
17. Кузнецов И.П., Мацкевич А.Г. Семантико-ориентированный лингвистический процессор для fвтоматической формализации автобиографических данных // Труды международной конференции покомпьютерной лингвистике и интеллектуальным технологиям "Диалог 2006", Бекасово, 2006, стр. 317-322.
18. Voss, S. and Joslyn C.A. Advanced Knowledge Integration in Assessing Terrorist Threats // LANL Technical Report LAUR 02-7867, 2002.
19. Сомин Н.В., Соловьева Н.С.., Шарнин М.М. Система морфологического анализа: опыт эксплуатации и модификации // Системы и средства информатики, Вып. 15 / Ин-т проблем информатики. – М.: Наука, 2005.
20. Gardner, J. R. and Z. L. Rendon, XSLT and XPATH: A Guide to XML Transformations, Prentice Hall, 2001.

Приложение 1.

Входной текст (на входе Semantix):

10:07 17.04.2003 США разбомбили в Ираке базы иранских террористов. Авиация США атаковала в Ираке базы иранской оппозиционной террористической группы "Моджахеддин хальк". Об этом в четверг со ссылкой на неназванные правительственные источники сообщило издание The New York Times. Как сообщили источники, один из ударов пришелся на штаб-квартиру оппозиционной группы в г. Ашрафе, в 95 км к северу от Багдада. Также нападению подверглись две другие базы. По оценкам американских властей, в состав оппозиционной группы "Моджахеддинхальк" входят несколько тысяч бойцов, большинство из которых базируется на иракской территории. //Reuters, Газета.Ru

XML-файл (на выходе Semantix):

```
<?xml version="1.0" encoding="windows-1251"?>
<DOCUMENT DOC_NUM="0">
  <OBJECT ID="1" TYPE="Place">
     <ATTR TYPE="STATE" VALUE="США"/>
     <SOURCE> США</SOURCE>
  </OBJECT>
  <OBJECT ID="2" TYPE="Terrorizm">
     <ARG CONST="ГОС-ВО"/>
     <ARG CONST="ИРАК"/>
```

```
    <ARG CONST=”БАЗА“/>
    <ARG CONST=”ИРАНСКИЙ“/>
    <ARG CONST=”ТЕРРОРИСТ“/>
    <SOURCE> Ираке базы иранских террористов</SOURCE>
  </OBJECT>
  <OBJECT ID=”3” TYPE=”Date”>
    <ATTR TYPE=”YEAR” VALUE=”2003”/>
    <ATTR TYPE=”MONTH” VALUE=”04”/>
    <ATTR TYPE=”DAY” VALUE=”17”/>
    <ATTR TYPE=”HOUR” VALUE=”10”/>
    <ATTR TYPE=”MINUTE” VALUE=”07”/>
    <SOURCE> 10 07 17. 04. 2003</SOURCE>
  </OBJECT>
  <OBJECT ID=”4” TYPE=”Military_Force”>
    <ARG CONST=”АВИАЦИЯ“/>
    <ARG CONST=”США“/>
    <SOURCE> Авиация США</SOURCE>
  </OBJECT>
  <OBJECT ID=”5” TYPE=”Terrorizm”>
    <ARG CONST=”ГОС-ВО“/>
    <ARG CONST=”ИРАК“/>
    <ARG CONST=”БАЗА“/>
    <ARG CONST=”ИРАНСКИЙ”/>
    <ARG CONST=”ОППОЗИЦИОННЫЙ”/>
    <ARG CONST=”ТЕРРОРИСТИЧЕСКИЙ”/>
    <ARG CONST=”ГРУППА“/>
    <ARG CONST=”МОДЖАХЕДДИН“/>
    <ARG CONST=”ХАЛЬК“/>
    <SOURCE> Ираке базы иранской оппозиционной террористической группы
Моджахеддин хальк</SOURCE>
  </OBJECT>
  <OBJECT ID=”6” TYPE=”Organization”>
    <ARG CONST=”NEW”/>
    <ARG CONST=”YORK”/>
    <ARG CONST=”TIMES”/>
    <SOURCE> New York Times</SOURCE>
  </OBJECT>
  <OBJECT ID=”7” TYPE=”Organization”>
    <ARG CONST=”ШТАБ“/>
    <ARG CONST=”КВАРТИРА”/>
    <ARG CONST=”ОППОЗИЦИОННЫЙ”/>
    <ARG CONST=”ГРУППА“/>
    <SOURCE> Штаб квартиру оппозиционной группы</SOURCE>
  </OBJECT>
  <OBJECT ID=”8” TYPE=”Address”>
    <ATTR TYPE=”CITY” VALUE=”АШРАФА“/>
    <ARG CONST=”95”/>
    <ARG CONST=”КМ.”/>
    <ARG CONST=”СЕВЕР“/>
    <ATTR TYPE=”CITY” VALUE=”БАГДАД“/>
    <SOURCE> Ашрафе, в 95 км к северу от Багдада</SOURCE>
  </OBJECT>
  <RELATION TYPE=”ИМЕТЬ“>
    <ARG REF=”7”/>
    <ARG REF=”8”/>
  </RELATION>
```

```
<OBJECT ID=”9” TYPE=”Position”>
   <ARG CONST=”АМЕРИКАНСКИЙ”/>
   <ARG CONST=”ВЛАСТЬ“/>
   <SOURCE> Американских властей</SOURCE>
</OBJECT>
<OBJECT ID=”10” TYPE=”Terrorizm”>
   <ARG CONST=”МОДЖАХЕДДИН“/>
   <ARG CONST=”ХАЛЬК“/>
   <SOURCE> Моджахеддин хальк</SOURCE>
</OBJECT>
<OBJECT ID=”11” TYPE=”Organization”>
   <ARG CONST=”REUTERS”/>
   <ARG CONST=”ГАЗЕТА“/>
   <SOURCE> Reuters, Газета</SOURCE>
</OBJECT>
<ACTION ID=”12” TYPE=”РАЗБОМБИТЬ“>
   <ARG REF=”1”/>
   <ARG CONST=”В“/>
   <ARG REF=”2”/>
</ACTION>
<RELATION TYPE=”КОГДА“>
   <ARG REF=”12”/>
   <ARG REF=”3”/>
</RELATION>
<ACTION ID=”13” TYPE=”АТАКОВАТЬ“>
   <ARG REF=”4”/>
   <ARG CONST=”В“/>
   <ARG REF=”5”/>
</ACTION>
<ACTION ID=”14” TYPE=”СООБЩИТЬ“>
   <ARG CONST=”НЕНАЗВАННЫЙ”/>
   <ARG CONST=”ПРАВИТЕЛЬСТВЕННЫЙ”/>
   <ARG CONST=”ИСТОЧНИК”/>
   <ARG CONST=”ИЗДАНИЕ“/>
   <ARG REF=”6”/>
</ACTION>
<ACTION ID=”15” TYPE=”УДАР“>
   <ARG CONST=”НА“/>
   <ARG REF=”7”/>
</ACTION>
<ACTION ID=”16” TYPE=”НАПАДЕНИЕ“>
   <ARG CONST=”2”/>
   <ARG CONST=”ДРУГОЙ“/>
   <ARG CONST=”БАЗА“/>
</ACTION>
<ACTION ID=”17” TYPE=”ВХОДИТЬ“>
   <ARG REF=”10”/>
   <ARG CONST=”НЕСКОЛЬКО“/>
   <ARG CONST=”1000”/>
   <ARG CONST=”БОЕЦ“/>
</ACTION>
<ACTION ID=”18” TYPE=”БАЗИРОВАТЬСЯ“>
   <ARG CONST=”НА“/>
   <ARG CONST=”ИРАКСКИЙ“/>
   <ARG CONST=”ТЕРРИТОРИЯ”/>
</ACTION>
```

```
  <SENTENCE>
     <ARG REF=”12”/>
<SOURCE>10:07 17.04.2003 США разбомбили в Ираке базы иранских террористов.
</SOURCE>
  </SENTENCE>
  <SENTENCE>
     <ARG REF=”13”/>
<SOURCE>Авиация США атаковала в Ираке базы иранской оппозиционной
террористической группы “Моджахеддин хальк”. </SOURCE>
  </SENTENCE>
  <SENTENCE>
     <ARG CONST=”ОБ“/>
     <ARG CONST=”ЧЕТВЕРГ“/>
     <ARG CONST=”СО“/>
     <ARG CONST=”ССЫЛКА“/>
     <ARG REF=”14”/>
<SOURCE>Об этом в четверг со ссылкой на неназванные правительственные
источники сообщило издание The New York Times. </SOURCE>
  </SENTENCE>
  <SENTENCE>
     <ARG CONST=”СООБЩИТЬ“/>
     <ARG CONST=”ИСТОЧНИК“/>
     <ARG CONST=”1”/>
     <ARG CONST=”ИЗ“/>
     <ARG REF=”15”/>
     <ARG REF=”8”/>
<SOURCE>Как сообщили источники, один из ударов пришелся на штаб-квартиру
оппозиционной группы в г.Ашрафе, в 95 км к северу от Багдада. </SOURCE>
  </SENTENCE>
  <SENTENCE>
     <ARG REF=”16”/>
<SOURCE>Также нападению подверглись две другие базы. </SOURCE>
  </SENTENCE>
  <SENTENCE>
     <ARG CONST=”ПО“/>
     <ARG CONST=”ОЦЕНКА“/>
     <ARG REF=”9”/>
     <ARG CONST=”СОСТАВ“/>
     <ARG CONST=”ОППОЗИЦИОННЫЙ”/>
     <ARG CONST=”ГРУППА“/>
     <ARG REF=”17”/>
     <ARG CONST=”БОЛЬШИНСТВО“/>
     <ARG CONST=”ИЗ“/>
     <ARG REF=”18”/>
<SOURCE>По оценкам американских властей, в состав оппозиционной группы
“Моджахеддин хальк” входят несколько тысяч бойцов, большинство из которых
базируется на иракской территории. </SOURCE>
  </SENTENCE>
  <SENTENCE>
     <ARG REF=”11”/>
<SOURCE>//Reuters, Газета.</SOURCE>
  </SENTENCE>
</DOCUMENT>
```