Средства настройки процессора Semantix на предметную область

Means for tuning of linguistic processor «Semantix» on subject field

Кузнецов И. П. (igor-kuz@mtu-net.ru)
Институт проблем информатики Российской академии наук

Ефимов Д. А. (d.efimov@synsys.ru)
ЗАО «Синергетические Системы»

Рассматривается семантико-ориентированный лингвистический процессор, осуществляющий автоматическую формализацию потоков текстов на естественном языке. В качестве исходного материала использован корпус текстов, связанный с описанием памятников. Из таких текстов выделяются информационные объекты: памятники, места их расположения, лица, их ролевые функции, события с указанием участия в них лиц и др. Рассматривается инструментальная среда, позволяющая быстро находить ошибки процессора и устранять их, подстраивая лингвистические знания.

Введение

С каждым годом возрастает совокупный объем цифровой информации. В основном, это тексты естественного языка (ЕЯ). Важная проблема — выбор из этих текстов информации, необходимой для профессиональной деятельности той или иной категории пользователей, и предоставление ее в форме, удобной для восприятия.

Следует учитывать, что большинство пользователей интересуются лишь конкретными вещами. Например, следователям важны фигуранты, их места жительства, телефоны и др. Специалистов по кадрам интересуют организации, где человек работал, кем и когда это было. Других интересуют памятные места, их местонахождение, кто автор, архитектор и т.д. Подобную информацию будем называть ***информационными объектами***, которые различаются по типам. Например, лица и фигуранты — это объекты одного типа, адреса — другого и т.д. Одно из важнейших направлений автоматической обработки потоков текстов — выявление информационных объектов и связей из текстов ЕЯ с ориентацией на определенную категорию пользователей. Это направление связано с формализацией текстов и относится к области «извлечение знаний» (Knowledge Extraction). При этом результаты должны быть представлены в формах, к которым привык пользователь (или же в формах, удобных для последующей обработки — поиска, экспертных оценок). Эта область в силу ее актуальности привлекает все больше исследователей, [4,5 и др.].

Отметим, что связи между объектами могут иметь высокую степень разнообразия. Например, памятники могут быть связаны не только с местом или лицами, которым они посвящены, но и с действиями или событиями — инициированием работ, проектированием, архитектурными работами, изготовлением отдельных компонент (постамента, фигур,...) и многое другое. Такие события привязаны к времени, месту, связаны с лицами, участвующие в создании памятников. Одни события могут быть составной частью других. Они могут быть связаны причино-следственными и временными отношениями. Таким образом, события — это тоже информационные объекты, связанные между собой и с другими информационными объектами. В ЕЯ такие связи выражаются с помощью глагольных форм, форм с отглагольными существительными, различными оборотами. Возникают сложные структуры.

Для представления подобных структур и их формализации были разработаны ***расширенные семантические сети*** (РСС). Для автоматического анализа текстов ЕЯ с их отображением на РСС в рамках проектов ИПИ РАН была разработана научная база для построения ***семантико-ориентированных лингвистических процессоров*** (ЛП): методики представления сложных видов знаний, ***инструментальная***

среда ДЕКЛ для обработки структур знаний, сетевые позиционные грамматики, онтологии, морфологический анализ на основе обобщенных окончаний, базы знаний (БЗ) на РСС, различные виды объектных поисков []. На этой основе разработан ряд прикладных систем [3,10,11]. Последний вариант ЛП, изготовленного совместно с ЗАО «Синергетические Системы» в виде модуля SDK, получил название Semantix и иллюстрировался на предыдущей конференции Диалог-08 [11]. Рассматривались различные аспекты организации подобных ЛП и особенности их работы.

Целью настоящей статьи является анализ важной компоненты семантико- ориентированных ЛП — средств настройки на предметную область. Здесь важную роль играют следующие факторы: как устроены лингвистические знания (ЛЗ), а также средства и методики их отладки или подстройки под предметную область пользователя. В качестве примера использован корпус текстов, связанный с описанием памятников. Данная предметная область достаточно интересная — со своими особенностями. В ней имеют место стандартные объекты (лица, даты, организации, профессии), а также объекты — памятники (компоненты их описания), места расположения, связанные с ними лица, события.

С точки зрения лингвистики область не является тривиальной. Пользователю важно знать, кто является инициатором памятника, архитектором, скульптором и др. Когда произошла закладка памятника, его установка, открытие и т.д. Чтобы выделить эти сведения, необходим глубинный анализ текста с выявлением имеющих место событий, для которых ищется привязка ко времени и месту.

1. Особенности семантико-ориентированных процессоров

Семантико-ориентированные ЛП, осуществляющие выявление информационных объектов и связей, основаны на правилах выделения компонент текста (слов, словосочетаний), из которых составляются информационные объекты. Такие правила в той или иной степени учитывают наличие ключевых слов, признаки слов (лексические, морфологические, семантические), взаимное расположение слов (их позиции), а также контекст.

В настоящее время развиваются два основных направления, связанных с построением семантико-ориентированных ЛП. Первое — когда правила «вшиваются» в программы. Создаются блоки, анализирующие слова, признаки, согласованность слов, наличие комбинаций цифр, знаков и др. Из таких блоков строятся правила. Каждое правило — это программа анализа с выделением объектов. Такие информационные объекты, как лица (ФИО), даты, адреса и организации с достаточной степенью надежности выделяются программными средствами. Анализ предложений сводится к выявлению наличия в них объектов, ключевых слов, значимых глаголов, например, *СОЗДАТЬ… ПРОЕКТ…*<лицо> и др. Лингвист задает правила анализа (модели), на основе которых из блоков строятся программы. При настройке на новые объекты лингвисту нужно разрабатывать новые правила, а программисту строить новые программы. Если лингвист чего-то не учел, то и программы будут давать ошибочные результаты. Нужно снова обращаться к лингвисту и т.д. В силу многообразия языковых конструкций, используемых при описании объектов (даже на корпусах текстов сравнительно небольших объемов), учесть все варианты представляется крайне трудной проблемой. Поэтому процесс настройки будет многоэтапным с определенной степенью сходимости.

В тоже время, программные средства постоянно совершенствуются (C#, .DOT и др.). Правила могут быть оформлены как программные объекты с экземплярами, учитывающими различные детали. В связи с этим процесс построения правил упрощается. Данный подход эффективно применяется в тех случаях, когда не требуется сложных видов анализа, например, связанных с выделением семантически связанных слов, событий с их атрибутами и др. Это направление развивается в ряде организаций, в том числе, в ЗАО «Синергетические Системы».

Второе направление, когда программа ***лингвистического процессора*** (ЛП) отделяется от ***лингвистических знаний*** (ЛЗ). Последние состоят из правил выделения объектов, включают в себя предметные словари, а также другие средства, определяющие всю процедуру анализа. ЛЗ имеют вид декларативных структур, которые легко менять и настраивать. В нашем случае роль таких структур выполняют фрагменты РСС [1,2]. Настройка ЛП осуществляется только за счет разработки ЛЗ, определяющих набор выделяемых объектов и связей. Задача ЛП — поддерживать ЛЗ, в том числе, процесс применения правил. При использовании подобных ЛП облегчается настройка на корпуса текстов, особенности предметной области. Корректировать ЛЗ может человек, обученный формализму РСС и знакомый с элементами математической лингвистики. Ему не нужно уметь программировать. Тогда возникает вариант, когда один человек может настраивать ЛП — находить ошибки и устранять их.

В данной статье речь будет идти о таких ЛП, к которым относятся процессоры системы «Аналитик» (с ее приложениями — системами «Криминал», «Обработчик резюме» и др.), а также процессор Semantix. Они все работают по одному принципу. Будем называть их **процессорами типа Semantix** (или просто Semantix). Отметим, что перечисленные системы отличаются только ЛЗ, которые определяют область приложений [3,6].

Основные компоненты семантико-ориентированного ЛП, основанного на ЛЗ:

1.1. Блок лексико-морфологического анализа

Выделяет из документа слова и предложения и выдает в виде семантической сети (***ПС-документа***), представляющей последовательность компонент (слов в нормальной форме, чисел, знаков) и их основные признаки, Использует набор предметных словарей (словарь стран, регионов России, имен, профессий и др.) для придания словам и словосочетаниям дополнительных семантических признаков [9].

1.2. Блок синтактико-семантического анализа

Путем анализа ПС-документа выделяет объекты и связи. На их основе строит другую семантическую сеть, представляющую семантическую структуру (***СС-документа***), называемую ***содержательным портретом*** [2,6,7]. Блок управляется ЛЗ, за счет которых обеспечивается: — Извлечение информационных объектов (лиц, организаций, событий, их места, …). — Выявление связей объектов. Например, как лица связаны с организациями, адресами и др. — Анализ глагольных форм, причастных и деепричастных оборотов с выявлением фактов участия объектов в тех или иных действиях. — Выявление связей действий с их местом или временем (где и когда имело данное действие или событие). — Анализ причино-следственных и временных связей между действиями и событиями.

1.3. Блок построения каталогов объектов

Выделяет из СС-документов объекты определенного типа, которые упорядочиваются по алфавиту и образуют каталог. Например, таким способом создаются каталоги лиц (их ФИО), дат, адресов и др. — только тех, которые встретились в документах.

1.4. База лингвистических знаний (ЛЗ)

Содержит правила анализа текста во внутреннем представлении (РСС). Они определяют работу ЛП.

Отметим, что с каждым годом растет количество теоретических работ в данной области (в том числе, см. последние сборники «Диалог»). Для многих из них доведение до конечного продукта может оказаться проблематичным в силу сложности возникающих проблем. В данной работе рассматривается работающий ЛП, отлаженный на различных предметных областях [11]. Процессоры такого типа регулярно демонстрировались на конференциях «Диалог» в течении 10 лет [2].

2. Проблемы настройки на предметную область

При настройке на предметную область возникают следующие трудности. Во-первых, при наличии разнотипных объектов требуются соответствующие правила их выделения. Качество ЛП определяется трудоемкостью построения таких правил. В процессоре Semantix выделение всех объектов (их может быть более 40 типов) и событий осуществляются правилами, которые конструктивно оформлены одинаковым образом, и соответственно, которые работают однотипным методикам. Поэтому трудоемкость построения не высокая. Важно, что изменяются только ЛЗ, но не программы.

Во-вторых, при настройке возникают частые случаи ***коллизии*** правил выделения: одни правила могут захватывать слова, которые относятся к другим объектам и которые должны обрабатываться другими правилами. В связи с этим правила должны иметь средства их быстрой подстройки, ограничивающие возможность применения. В процессоре Semantix такая подстройка осуществляется за счет изменения списков, задающих допустимые признаки слов, стоящих на тех или иных позициях. В общем случае такие списки организованы в виде И-ИЛИ графов.

В-третьих, важный фактор — это **избирательность правил** и процедур идентификации: ***коэффициент шумов и потерь***. Под шумами понимается наличие лишних слов в объектах. Потери — это когда объект не выявлен или выявлен частично (в тексте есть слова, которые не вошли в объект). В процессоре Semantix правила (составляющие ЛЗ) имеют все средства для повышения степени избирательности правил и минимизацию шумов и потерь при большом количестве выделяемых объектов. С помощью ЛЗ обеспечивается настройка на особенности языка — на типовые конструкции и формы языка с учетом признаков, которые даются словам. Имеются все необходимые удобства в плане создания и корректировки правил.

В-четвертых, определенные трудности вызывает выделение связей. Это не только глубинный анализ глагольных и других форм. Многие связи даются с помощью анафорических ссылок, а также по умолчанию. Требуется организация сложного процесса их поиска. Такие процессы организуются, чтобы связать лицо с его местом проживания или местом работы, идентифицировать слова (*ПАМЯТНИК, ФИГУРА*,..) с объектами (типа *памятник*) и т.д. Эти слова и подразумеваемые объекты могут стоять в тексте на значительном расстоянии. Важно не захватить посторонний объект. В процессоре Semantix для этой цели используются специальные фильтры.

В-пятых, для настройки ЛП на корпуса текстов необходим специальный **комплекс инструментальных средств**, обеспечивающих следующие функции:

- последовательную обработку множества документов (корпуса текстов) с формализацией каждого из них, формированием СС-документов и построением общей БЗ;
- формирование списков выделенных объектов (для каждого типа объектов свой список), осуществляемое в процессе обработки множества документов; такие списки будем называть ***каталогами***;
- возможность выделения в каталоге любого объекта с быстрым поиском документов, из которых выделен данный объект;
- подача на вход ЛП найденного документа с анализом процесса его формализации (формирования СС-документа);
- визуализацию процесса применения правил и осуществляемых ими преобразований;
- трассировку работы каждого правила с указанием, в какой последовательности захватываются слова (благодаря каким признакам), а также, почему и на каком слове процесс применения правила закончился;
- выдачу в одно окно СС-документа и сам документ для их сравнения;
- обращение к ЛЗ с выбором любого правила и его изменением.

Эти средства позволяют быстро находить ошибки в работе ЛП и корректировать ЛЗ. Методика достаточно проста. Берется корпус текстов и включается в последовательную обработку с формированием общей БЗ и каталогов объектов. Они просматриваются. С их помощью легко находить объекты с ***шумами*** (лишними словами). Во многих случаях сразу видны ***потери*** — если по смыслу объект не соответствует своему статусу. Труднее находить потери, когда объект имеется в документах, но не найден. Тогда нужно просматривать каталог «лишних» слов и их комбинаций (которые не вошли ни в какие объекты). Или же просматривать СС-документов и сравнивать их с самими документами. Конечно, идеальный вариант, когда кто-то (лингвист) выделяет из корпуса текстов объекты определенного типа, и они сравниваются с построенным каталогом. Но этот вариант крайне трудоемок, когда имеют место корпуса текстов большого объема, которые постоянно обновляются.

Следует отметить, что подобные инструментальные средства отсутствуют в процессоре Semantix, изготовленного в виде модуля SDK (там выключена интерфейсная компонента). Но они имеются в системах «Аналитик» ("Криминал»), где реализованы различные виды объектных поисков, ответ на запросы в свободной форме, развита интерфейсная компонента. Поэтому процесс разработки и отладки ЛЗ для новой предметной области идет в рамках этих систем. Отлаженные ЛЗ переносятся в модуль SDK.

3. Выделяемые объекты и связи

Набор выделяемых объектов и связей определяется задачами пользователя. В рамках выполнения плановых работ была взята предметная область, связанная с описанием памятников. Работа в этой области явилась еще одним примером достаточной универсальности процессора Semantix. Корпус текста имеет вид:

6. *Памятник руководителю первой русской кругосветной экспедиции Ивану Фёдоровичу Крузенштерну находится на набережной Лейтенанта Шмидта. Решение об его установке было принято в 1869 году, в канун столетия со дня рождения адмирала. Памятник находится напротив здания Морского кадетского корпуса. Торжественная закладка памятника состоялась 8 ноября 1870 года, в день 100-летнего юбилея Крузенштерна. Фигура из бронзы была отлита в декабре 1872 года по модели И. Н. Шредера. Гранитный постамент спроектировал архитектор И. А. Монигетти. Открытие памятника состоялось 6 ноября 1873 года.*

7. *23 мая 1909 года в центре Знаменской площади был открыт конный памятник Александру III. Автор памятника Паоло Трубецкой выполнил его откровенно как карикатуру, что вызвало довольно сильный скандал. В 1937 году памятник перенесли во двор Михайловского дворца. Перенос памятника объяснили тем, что он якобы мешал трамвайному движению, хотя к тому времени трамвай по Невскому проспекту ходил уже около трёх десятилетий.*

8. *Памятник Екатерине II создан в 1862–1873 годах по проекту ...*

В качестве основных типов информационных объектов и связей были выбраны следующие:

- памятники (монументы, скульптуры и т.д.);
- лица (кому посвящен памятник, кто участвовал в его создании и др.);
- ролевые функции или профессии (скульптор, архитектор, дизайнер, зодчий, ...) ;
- места расположения памятников;
- события с указанием участия в них информационных объектов («памятник создан ...», «работа выполнена ...»);
- даты, время, интервалы времени; — организации (связанные с созданием памятников);
- связи между различными типами информационных объектов (время и место событий, ролевые функции лиц и др.).

Отметим, что выявление ролевых функций лиц и связей требует анализа различных форм ЕЯ (с однородными членами: отглагольными существитель-

ными, причасными: деепрчаастными оборотами и др.) для выделения событий с их привязкой к времени и месту. Требуется сложный (многоуровневый) лингвистический анализ с учетом различных признаков, в том числе семантических.

4. Правила выделения объектов

Правила, осуществляющие выделение лиц, дат, интервалов времени, профессий, организаций, событий и связей были взяты из предметных областей, на которые уже был настроен процессор Semantix. Они отлаживались на текстах — «Документы о терроризме», «Автобиографии», «Сводки происшествий» и др.

Новые информационные объекты — это памятники и (в значительной степени) места их расположения. Ранее не встречались описания типа: *возле Каменного моста, напротив здания Моссовета, на площади перед Михайловским замком* и др. Например, в сводках происшествий таких описаний не встречается — всегда фигурируют названия мест.

Были разработаны правила выделения памятников (их несколько). Правила имеют левую часть (условие применения) и правую (действия). К примеру, одно из них выглядит следующим образом:

MUSTBE(MONUM~2,1)
STR_OR(*БЮСТ,СТАТУЯ,ФИГУРА,СОБОР,ЦЕРКОВЬ,ХРАМ*/1+)
STR_OR(*WORK_K,NAT_K,ИМПЕРАТОР,ЦЕСАРЕВИЧ,ГРАФ,КНЯЗЬ,ГЕРЦОГ,..*/2+)
STR_OR(*КОГО,КВЧ,MONUM_K,FIO,FAM*/3+)
CONTEXT(1-,2-,3-/MONUM~2)
P_P(MONUM~2,MON~2L)
MONUM_(1,2,3/MON~2L) MON~2L(*MONUM*,ADD_)
MAYBE(MONUM~2,2)

Данное правило записано в формализме РСС и означает следующее. Вызов правила — по его идентификатору MONUM~2. Фрагмент P_P(MONUM~2,MON~2L) разделяет левую и правую части, т.е. CONTEXT(1-,2-,3-/MONUM~2), который задает условие применения, и MONUM_(1,2,3/MON~2L) — что формировать.

Применять правило нужно с 1-й позиции — MUSTBE(MONUM~2,1). Нужно искать ключевые слова, отмеченные 1+ , т.е. *БЮСТ, СТАТУЯ*,... На следующей позиции должно быть одно из слов списка 2+. Это может быть слово с признаком *WORK_K* (т.е. входить в словарь професссий) или с признаком *NAT_K* (словарь национальностей), или же одно из перечисленных далее слов. Фрагмент MAYBE(MONUM~2,2) указывает, что эта позиция факультативная — перечисленных слов может не быть в тексте).

На следующей позиции должно быть одно из слов списка 3+. Это может быть слово с признаком *КОГО* (род. падеж), или слово в кавычках, или слово с признаком *MONUM_K* (словарь памятников с уникальными названиями), или фамилия, или лицо (ФИО). Если условие выполняется, то правило будет применимым. Формируется объект MONUM_(1,2,3) с признаком *MONUM*. Аргументами являются слова (или объекты), которые оказались на позициях 1,2,3. Сформированный объект замещает эти слова и занимает свою позицию: три позиции замещаются на одну.

Правило будет применяться, когда в тексте встречаются описания типа *статуя императора Николая I, Фигура Петра Великого* и т.д.

Для уникальных памятников, которые невозможно выделить через ключевые слова, создан предметный словарь MONUM_K.SLV, фрагмент которого имеет вид:

. . .
Александровская колонна
«Железный Феликс»
«Лысый Камень»
Вандомская колонна
Воин-освободитель
Демидовский столп
<Медный Всадник>
Миноносец «Стерегущий»
Родина-мать
Собор Парижской богоматери
Соловецкий камень
«Шалаш»
«Царь-плотник»
. . .

Если в тексте встретилось одно из этих слов (или словосочетаний), то ему (им) присваивается признак *MONUM_K*, который учитывается правилами из ЛЗ. Следует отметить, что словарь *MONUM_K* сравнительно небольшой, как и набор ключевых слов. Основная часть описания памятника выделяется из текста ЕЯ. Это слова, составляющие окрестность ключевых слов , например, ФИО, слова в родительном падеже и др. (см. правило MONUM~2).

В результате применения правил формируется СС-документа (содержательный портрет), где все слова приведены в нормальную форму, а объекты и связи представлены в виде фрагментов РСС. Например, для 1-го документа (из корпуса текстов) он будет иметь вид;

ДОК_(1,MONUM.TXT,"ПАМЯТНИКИ;")
FIO(КРУЗЕНШТЕРН,ИВАН,ФЕДОРОВИЧ," "/1+)
РАБ_(1-,РУКОВОДИТЕЛЬ,ПЕРВЫЙ,РУССКАЯ,КРУГОСВЕТНЫЙ,ЭКСПЕДИЦИЯ/2+)
MONUMENT_(ПАМЯТНИК,2-/3+)
НАХОДИТЬСЯ(3-/4+)

АДР_(НАБ.,ЛЕЙТЕНАНТ,ШМИДТ/5+)
Где(4-,5-)
РЕШЕНИЕ(О,3-/6+)
ПРИНЯТЬ(6-,УСТАНОВКА/7+)
ДАТА_(1869,ГОД/8+)
Когда(7-,8-)
ОРГ_(МОРСКОЙ,КАДЕТСКИЙ,КОРПУС/9+)
PLACE_(НАПРОТИВ,ЗДАНИЕ,9-/10+)
Где(4-,10-)
. . .

Первый фрагмент говорит, документ взят из файла MONUM.TXT и имеет номер 1. Последующие три фрагмента представляют *«Памятник руководителю первой русской кругосветной экспедиции Ивану Фёдоровичу Крузенштерну»*. Следующие три фрагмента — *"он находится на набережной Лейтенанта Шмидта»* и т.д. Коды 1+ и 1- (2+ и 2- и т.д.) обозначают один и тот же объект — лицо FIO (соответственно, профессию — РАБ_). Более подробное описание того, как устроена СС-документа, см. в [2,8,9]. СС-документов составляют базу знаний и служат для решения задач. В частности, СС-документов являются исходным материалом для автоматического порождения различных сведений — кто автор, архитектор, когда памятник установлен, открыт и т.д. Это делается с помощью экспертных программ на языке ДЕКЛ, который создан для обработки структур знаний на РСС [2].

5. Каталоги объектов

Как уже говорилось, отладка правил и ЛЗ ведется в рамках систем «Аналитик» («Криминал»). Они обеспечивают последовательную обработку множества документов (корпуса текстов из заданного файла) с формализацией каждого из них, формированием СС-документов и автоматическим построением общей БЗ и каталогов объектов. Такие каталоги — это списки выделенных объектов, где слова представлены в нормальной форме (необходимо для поиска). Например, каталог выделенных памятников (когда обработано 30 документов) имеет вид:

АЛЕКСАНДРОВСКАЯ КОЛОННА
АЛЛЕГОРИЧЕСКИЙ ЖЕНСКИЙ ФИГУРА МУДРОСТЬ
В ПАМЯТЬ ЖЕРТВА РЕПРЕССИЯ ПОЛИТИЧЕСКИЙ
В ПАМЯТЬ О ПРИБЫТИЕ ЛЕНИН
ВАНДОМСКАЯ КОЛОННА
СКУЛЬПТУРА ВЕРБЛЮД
ГРАНИТНЫЙ ПОСТАМЕНТ
ЕКАТЕРИНИНСКИЙ ДВОРЕЦ
КОННЫЙ ПАМЯТНИК АЛЕКСАНДР III
МИНОНОСЕЦ СТЕРЕГУЩИЙ
МИХАЙЛОВСКИЙ ДВОРЕЦ
МИХАЙЛОВСКИЙ ЗАМОК
МОНУМЕНТАЛЬНЫЙ ПАМЯТНИК ЛЕРМОНТОВ М. Ю.
ОБЕЛИСК ГОРОДУ ГЕРОЮ ЛЕНИНГРАДУ
ПАМЯТНИК-БЮСТ ЗНАМЕНИТЫЙ АРХИТЕКТОР XVIII В. РАСТРЕЛЛИ ФРАНЧЕСКО Б
ПАМЯТНИК-БЮСТ МУСОРГСКИЙ М. П.
ПАМЯТНИК-БЮСТ НЕКРАСОВ Н. А.
ПАМЯТНИК-БЮСТ ПРЖЕВАЛЬСКИЙ Н. М.
ПАМЯТНИК-БЮСТ СЕМЕНОВ-ТЯН-ШАНСКИЙ П. П.
ПАМЯТНИК АДМИРАЛ КРУЗЕНШТЕРН И. Ф.
ПАМЯТНИК АЛЕКСАНДР I
ПАМЯТНИК ГОРЬКИЙ А. М.
ПАМЯТНИК ЕКАТЕРИНА II
ПАМЯТНИК ИВАН ФЕДОРОВИЧ
ПАМЯТНИК КРЫЛОВ ИВАН АНДРЕЕВИЧ
ПАМЯТНИК ЛОМОНОСОВ М. В.
ПАМЯТНИК МЕДНЫЙ ВСАДНИК ПЕТР I
ПАМЯТНИК МИНОНОСЕЦ СТЕРЕГУЩИЙ
ПАМЯТНИК НАЗЫВАТЬСЯ ГЕРОЙ КРАСНОДОНА
ПАМЯТНИК НИКОЛАЙ I
ПАМЯТНИК РУКОВОДИТЕЛЬ ПЕРВАЯ РУССКАЯ КРУГОСВЕТНАЯ ЭКСПЕДИЦИЯ ..
. . .

Каталог выделенных мест:
АЛЕКСАНДРОВСКИЙ ПАРК ПРОСП. КАМЕННООСТРОВСКОЙ
БОЛЬШОЙ ПРОСП. САМПСОНИЕВСКИЙ
В ЦЕНТР ПЛ. ДВОРЦОВЫЙ
В ЦЕНТР ПЛ. ЗНАМЕНСКАЯ
В ЦЕНТР ПЛ. СЕНАТСКИЙ
*ВЕРЕБЬИНСКИЙ МОСТ ЖЕЛЕЗНЫЙ ДОРОГА ГОРОД САНКТ-ПЕТЕРБУРГ — МОСКВА**
ДВОР МИХАЙЛОВСКИЙ ДВОРЕЦ
ИОАННОВСКИЙ МОСТ
НАБ. ЛЕЙТЕНАНТ ШМИДТ
НАБ. РЕКА НЕВА НАПРОТИВ МОРСКОЙ КАДЕТСКИЙ КОРПУС
НАБ. РЕКА ФОНТАНКА
НАБ. РЕКА ФОНТАНКА СКВЕР ЛОМОНОСОВСКИЙ
НАПРОТИВ ЗДАНИЕ МОРСКОЙ КАДЕТСКИЙ КОРПУС
ПАРК ИМЕНИ 30 ЛЕТИЯ ВЛКСМ
ПЕРЕД ЗДАНИЕ УЧИЛИЩЕ ПРОСП. ЛЕРМОНТОВСКОЙ
ПЕТЕРГОФСКИЙ ПАРК АЛЕКСАНДРИЯ
ПЛ. ОСТРОВСКИЙ СКВЕР ПЕРЕД АЛЕКСАНДРИНСКИЙ ТЕАТР
ПРОСП. КАРЛ МАРКС
САД ПРУДОК НА УГОЛ УЛ. БАССЕЙНОЙ
САНКТ ПЕТЕРБУРГ
СКВЕР ЕКАТЕРИНА ПЕРЕД ТЕАТР
СКВЕР ПЕРЕД АЛЕКСАНДРИЙСКИЙ ТЕАТР
СКВЕР ПЛ. ТРОИЦКАЯ
УГОЛ КРОНВЕРКСКИЙ ПРОСП. КАМЕННООСТРОВСКИЙ НА НЕБОЛЬШОЙ ..
ЦЕНТР ГЛАВНЫЙ ПЛ. ГОРОД САНКТ ПЕТЕРБУРГ

Выявление шумов и потерь сводится к просмотру таких каталогов с поиском неполных или бессмысленных описаний. Выбрав любую строчку и нажав ENTER, будет найден документ, из которого выделен соответствующий объект. Этот документ подается на вход ЛП для повторного анализа.

Еще раз отметим, что такие каталоги строятся автоматически в процессе анализа корпусов текстов с выделением объектов. Сформированный каталог можно поместить в соответствующий предметный словарь, расширив его. Однако, это не приведет к существенному улучшению работы ЛП, так как имеющиеся в каталоге объекты и так устойчиво выделяются ЛП.

6. Просмотр процесса применения правил

Когда найден документ, из которого не правильно выделен объект, нужно найти соответствующее правило и скорректировать его. Для этого служит **режим просмотра** процесса анализа документа, где для каждого правила указывается, какие оно осуществило преобразования. Процесс визуализации изображен на рис.1.

На рис.1 показано следующее. Правило FF~5 применилось и выделило лицо *МОНИГЕТТИ И. А.* Следующее правило объединило фамилию *КРУЗЕНШТЕРН* с выделенным лицом — *КРУЗЕНШТЕРН ИВАН ФЕДОРОВИЧ*. Слово РАБ_ указывает на профессию, а OBJ_ — на памятники. Правила GG~1 и GG~2 выделяют группы согласованных слов — генетивные цепочки. Правила MONUM~1, MONUM~4 выделяют группы слов — описания памятников. В режиме просмотра легко увидеть, что сделало каждое правило и где имела место ошибка. Более того, все правила имеют встроенные **механизмы трассировки**. Их можно активизировать для каждого интересующего правила. Трассировка визуализирует процесс, выводя на экран, в какой последовательности захватываются слова и почему правило оказалось не применимым [11,12].

7. Выдача результатов

Главным в работе ЛП является формирование СС-документов (см. п. 4), которые пользователь не должен видеть. Но на основе СС-документов с помощью достаточно простых ДЕКЛ-программ могут строиться различные формы или описания, необходимые для пользователя. Например, это может быть XML-файл, где для каждого объекта дается набор

.....
ФИО: *КРУЗЕНШТЕРН ИВАН ФЕДОРОВИЧ* << FF~4
ФИО: *МОНИГЕТТИ И. А.* << FF~5
ФИО: *КРУЗЕНШТЕРН ИВАН ФЕДОРОВИЧ & ФИО: КРУЗЕНШТЕРН* — одно лицо << FIO_UN
+++ Уровень +++ LEVEL_T3
+++ Уровень +++ LEVEL_T4
WORD_C: *КРУГОСВЕТНЫЙ ЭКСПЕДИЦИЯ* << GG~1
WORD_C: *РУССКАЯ КРУГОСВЕТНЫЙ ЭКСПЕДИЦИЯ* << GG~1
WORD_C: *ПЕРВЫЙ РУССКАЯ КРУГОСВЕТНЫЙ ЭКСПЕДИЦИЯ* << GG~1
РАБ_: *РУКОВОДИТЕЛЬ ПЕРВЫЙ РУССКАЯ КРУГОСВЕТНЫЙ ЭКСПЕДИЦИЯ* << GG~2
WORD_C: *ТОРЖЕСТВЕННЫЙ ЗАКЛАДКА* << GG~1
WORD_C: *КАНУН СТОЛЕТИЕ* << GG~2
WORD_C: *100-ЛЕТНИЙ ЮБИЛЕЙ* << GG~2
РАБ_: *ШРЕДЕР И. Н. МОДЕЛЬ* << WORK~1A
РАБ_: *МОНИГЕТТИ И. А. АРХИТЕКТОР* << WORK~1A
РАБ_: *КРУЗЕНШТЕРН ИВАН ФЕДОРОВИЧ РУКОВОДИТЕЛЬ ПЕРВЫЙ РУССКАЯ КРУГОСВЕТНЫЙ ЭКСПЕДИЦИЯ* << WORK~1A
OBJ_: *ПАМЯТНИК КРУЗЕНШТЕРН ИВАН ФЕДОРОВИЧ РУКОВОДИТЕЛЬ ПЕРВЫЙ РУССКАЯ КРУГОСВЕТНЫЙ ЭКСПЕДИЦИЯ* << MONUM~1
OBJ_: *ФИГУРА ИЗ БРОНЗЫ* << MONUM~4
OBJ_: *ГРАНИТНЫЙ ПОСТАМЕНТ* << MONUM~4
+++ Уровень +++ LEVEL_T5
ОН — это *ПАМЯТНИК КРУЗЕНШТЕРН ИВАН ФЕДОРОВИЧ* ... << ID_2MM
ПАМЯТНИК — это *ПАМЯТНИК КРУЗЕНШТЕРН ИВАН ФЕДОРОВИЧ* ... << ID_2W
.....

ОТЛИТЬ: ФИГУРА ИЗ БРОНЗЫ ПО ШРЕДЕР И.Н. МОДЕЛЬ << VV~1

Когда: ДАТА_: *ДЕКАБРЬ 1872* . . .

Рис. 1. Визуализация процесса анализа

составляющих его слов в нормальной форме (для поиска) и его описание, взятое из текста [11] . Описания выделенных объектов могут служить основой для заполнения БД или полей какого-либо сайта (формы), как это сделано в системе формализации резюме.

Ниже приведен еще один пример, когда для лиц даются из ролевые функции, а для времени и места указываются события, к которым они относятся.

```
<DOCUMENT DOC_NUM="0">
    <OBJECT ID="1» TYPE="FIO">
        <ARG TYPE=" -- Посвящен:"/>
        <SOURCE> Ивану Федоровичу Крузенштерну</SOURCE>
    </OBJECT>
    <OBJECT ID="2» TYPE="USER_OBJECT">
        <SOURCE> Памятник руководителю первой русской кругосветной экспедиции Ивану
                 Федоровичу Крузенштерну</SOURCE>
    </OBJECT>
    <OBJECT ID="3» TYPE="ADDRESS">
        <ARG TYPE=" -- ГДЕ ... находится:"/>
        <SOURCE> на набережной Лейтенанта Шмидта</SOURCE>
    </OBJECT>
    <OBJECT ID="4» TYPE="DATE">
        <ARG TYPE=" -- КОГДА ... решение об его установке было принято:"/>
        <SOURCE> в 1869 году</SOURCE>
    </OBJECT>
    <OBJECT ID="5» TYPE="PLACE">
        <ARG TYPE=" -- ГДЕ ... находится:"/>
        <SOURCE> Напротив здания Морского кадетского корпуса </SOURCE>
    </OBJECT>
    <OBJECT ID="6» TYPE="DATE">
        <ARG TYPE=" -- КОГДА ... закладка памятника состоялась:"/>
        <SOURCE> 8 ноября 1870</SOURCE>
    </OBJECT>
    <OBJECT ID="7» TYPE="USER_OBJECT">
        <SOURCE> Фигура из бронзы</SOURCE>
    </OBJECT>
    <OBJECT ID="8» TYPE="FIO">
        <ARG TYPE=" -- Создатель модели:"/>
    <SOURCE> И.Н. Шредера</SOURCE>
    ...
```

Формирование такого XML-файла осуществляется с помощью ДЕКЛ-программ, которые анализируют СС-документа и присваивают новые свойства объектам. Отметим, что отдельные компоненты описания взяты из текста и поэтому могут быть в различных падежных формах, например, *И. Н. Шредера*. В настоящее время разработана программа их корректировки, которая в данной работе не использована.

Заключение

В настоящее время лингвистический процессор Semantix настроен на автоматическую обработку потоков текстов на естественном языке (ЕЯ), представляющих собой: резюме на русском и английском языке, сообщения СМИ (о терактах), тексты описания достопримечательностей (памятников), сводки происшествий, справки по уголовным делам. Процессор может быть использован для обработки архивных и информационно-рекламных материалов, почтовых сообщений и т. д. Достоинства этого процессора — высокая избирательность при выделении объектов и связей, наличие правил глубинного анализа с выделением событий и их привязкой к времени и месту, а также наличие средств быстрой настройки на новую предметную область. Как показывает опыт, время такой настройки исчисляется не годами, а неделями, месяцами: зависит от количества и сложности новых объектов и допустимыми коэффициентами шумов-потерь.

ДЕМО-версия процессора Semantix — http://www.semantix4you.com

Литература

1. *Кузнецов И. П.* Семантические представления // М. Наука. 1986 г. 290 с.
2. *Кузнецов И. П., Мацкевич А. Г.* Семантико-ориентированные системы на основе Баз Знаний.// Монография, МТУСИ. М.: 2007. 173 с.
3. *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. Тарусса 1999.
4. *Cunningham, H.* Automatic Information Extraction. // Encyclopedia of Language and Linguistics, 2cnd ed. Elsevier, 2005.
5. *Han J. and Kamber, M.* Data Mining: Concepts and Techniques // Morgan Kaufmann, 2006.
6. *Igor Kuznetsov, Elena Kozerenko.* 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.
7. *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.
8. *Kuznetsov I. P., Kozerenko E. B.* Linguistic Processor «Semantix» for Knowledge extraction from natural texts in Russia and English. Proceeding of International Conference on Machine Learning, ISAT-2008. 14–18 July, 2008 Las Vegas, USA// CSREA Press, 2008, p.835–841.
9. *Кузнецов И. П., Сомин Н. В.* Средства настройки семантико-ориентированного лингвистического процессора на выделение и поиск объектов. Сб. ИПИ РАН, Вып.18. 2008 г., стр. 119–143 .
10. *Кузнецов И. П.* Объектно-ориентированная система, основанная на знаниях в виде XML-представлений.// Сб. ИПИ РАН, Вып.18.М.: 2008. С. 96–118.
11. *Кузнецов И. П., Ефимов Д. А.* Особенности извлечения знаний семантико-ориентированным лингвистическим процессором Semantix.// Сб. Компьютерная лингвистика и интеллектуальные технологии. Выпуск 7 (14). По материалам конференции «Диалог 008»..РГГУ, М.:2008., С. 281–291.