Метаданные в системе управления многоязычной лингвистической базой знаний

© Н.В. Лунева

Институт проблем информатики РАН
nl2@mail.ru

Аннотация

В работе рассматривается использование метаданных в системе управления функционированием многоязычной лингвистической базы знаний, их роль в процессе взаимодействия функциональных подсистем и блоков программного комплекса базы знаний. Метаданные применяются для описания обрабатываемых и служебных данных системы, их текущего состояния и состояния самой системы, а также для передачи информации между функциональными подсистемами и блоками многоязычной лингвистической базы знаний и управления их работой.

1 Введение

В настоящее время метаданные являются широко распространенным инструментом, используемым для:

- обмена сведениями между различными инструментами и подсистемами информационной системы,
- хранения информации о внутренней структуре, форматах и содержании информационных ресурсов системы,
- интеграции разнородных данных системы в единое информационное пространство,
- описания самой системы.

Многоязычная лингвистическая база знаний создается в рамках проекта «Интертекст» на основе параллельных научных и патентных текстов и метаязыковых лингвистических представлений и предназначена для отладки семантико-синтаксических представлений в лингвистических процессорах систем машинного перевода и обработки текстовых знаний. Концептуальные и лингвистические основы данного проекта изложены в работах Е.Б. Козеренко [1-6, 9-11], общее описание архитектуры и метаданных, интерфейса и функциональных компонент многоязычной лингвистической базы знаний рассмотрены в работе [7].

В основе многоязычной лингвистической базы знаний лежит разработка системы правил фразовых структур, функциональные значения языковых единиц закодированы как метки фразовых структур, и типы атрибутов-значений определяются функционально-категориальной семантикой. Множество языковых структур, представленных в виде синтактико-семантических комплексов, выстраиваются в иерархию правил. Отношения зависимости реализуются через механизм головных вершин фразовых структур, а сами фразовые структуры задают линейные последовательности языковых объектов [1, 10].

Программный комплекс лингвистической базы знаний обеспечивает ввод обрабатываемого текста и его разметку, распознавание функционально-семантических структур введенного текста, построение на их основе набора соответствующих структур для результирующих текстов и формирование самих текстов, а также подбор примеров из архивов базы знаний. Кроме того, обеспечивается управление функционированием системы и ее настройками, служебными и пользовательскими архивами, использованием словарей [7].

Данные, используемые лингвистической базой знаний, представляют собой набор разнородных массивов информации, хранящейся в различных форматах – коллекции параллельных текстов научных и патентных документов, коллекции «переводческой памяти», комплексы фразовых структур, словари, служебные файлы и пользовательские архивы. Метаданные столь разнородной информации образуют комплекс, каждая часть которого описывает свой специфический класс данных системы.

Метаданные в многоязычной лингвистической базе знаний используются как для описания массивов данных и коллекций системы, так и в сфере управления функционированием блоков и подсистем базы знаний, а также для организации связи между ними в процессе обработки текста (распознавания и трансфера) и взаимодействия с пользователем. Метаданные, описывающие текущее состояние программного комплекса в процессе выполнения задания (метаописание), формируются в самом начале сеанса работы пользователя на основе шаблона либо загружаются из вызванного при входе в систе-

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

му предыдущего сеанса работы. В процессе работы метаописание модифицируется, чтобы отображать текущее состояние системы и проекта, что позволяет при необходимости приостановить заданный процесс обработки текста и впоследствии продолжить его с того же самого места, а также изменять текущие настройки и параметры среды.

2 Метаописание в многоязычной лингвистической базе знаний

Наряду с использованием метаданных для описания архивов и коллекций многоязычной лингвистической базы знаний и их содержимого, метаданные применяются для организации взаимодействия функциональных блоков и подсистем между собой. Метаданные используются для управления работой программного комплекса многоязычной лингвистической базы знаний, процессами распознавания введенного текста, построения функционально-семантических структур, трансфера, подбора примеров и формирования целевого текста. Метаданные обеспечивают обмен информацией о состоянии и выполняемых действиях, требуемых для работы внутренних данных между программными элементами базы знаний, и управление коллекциями данных лингвистической базы знаний. Кроме того, метаданные хранят настройки программной среды на конкретную задачу и пользователя, обеспечивая взаимодействие с ним.

Набор метаданных, хранящий описание текущего состояния многоязычной лингвистической базы знаний в процессе выполнения конкретной задачи, естественно назвать метаописанием. Каждое метаописание уникально для пользователя, решаемой им задачи и стадии выполняемого процесса. В процессе работы программного комплекса метаописание постоянно модифицируется, чтобы отображать текущее состояние системы и задачи, что позволяет при необходимости приостановить заданный процесс обработки текста и впоследствии продолжить его с того же самого места, а также изменять текущие настройки и параметры среды. Принципиальная схема взаимодействия функциональных блоков и подсистем многоязычной лингвистической базы знаний посредством метаописания приведена на рис. 1.

При самом первом входе пользователя в среду базы знаний формируется исходное метаописание на основе стандартного шаблона и конфигурации. Настройки, выбираемые пользователем для удобства работы в системе и в целях обработки конкретного текста либо другого задания, отражаются в метаописании, как и задаваемые пользователем задачи и их состояние.

Когда пользователь приостанавливает свою работу и выходит из системы, метаописание сохраняется, и при следующем входе пользователь имеет возможность продолжить приостановленный процесс с той самой точки, где работа была прервана. С другой стороны, настройки среды «под себя», сделанные пользователем, сохраняются в файле конфигурации, и при его следующем входе может быть сформировано исходное метаописание уже на их основе.

Функционирование многоязычной лингвистической базы знаний основывается на взаимодействии между собой ряда подсистем. При передаче обрабатываемого текста от одной подсистемы или блока

Рис. 1. Схема взаимодействия функциональных блоков и подсистем многоязычной лингвистической базы знаний через метаописание текущего состояния. Цифрами обозначены процессы:

1 – формирования метаописания задачи,

2 – управления функционированием,

3 – модификации текущего метаописания,

4 – настройки и управления системой, а также модификации метаописания.

другим вместе с ним необходимо передавать информацию о текущем состоянии обрабатываемых данных, среде обработки, требуемых или ожидаемых далее результатах и т.п. Передача такой информации от одной подсистемы другим, как и информации о том, какие именно подсистемы и службы базы знаний должны далее отработать над текущей задачей, выполняется через метаописание.

Подсистема обработки входного текста обеспечивает ввод исходного текста и его характеристик в пользовательскую среду базы. Далее выполняется его подготовка для подачи в подсистему распознавания функционально-семантических структур – распознавание внутренней структуры текста и его фрагментация на логико-структурные блоки – фразы и фрагменты фраз.

Подсистема распознавания структуры входного фрагмента обеспечивает предварительное распознавание словарного состава заданного фрагмента текста и его морфологическую разметку для дальнейшего семантико-синтаксический анализа по правилам комплекса фразовых структур входного языка, построенных на основе новой когнитивной трансферной грамматики, впервые представленной в работах [3,9], и специализированных словарей входного языка, и построение его формальной семантической структуры. Результаты анализа отображаются в заданной пользователем нотации для оценки соответствия построенных структур исходному тек-

сту. Предусматривается возможность коррекции получаемых результатов и выбор структуры для дальнейшей работы системы, наиболее релевантной тексту.

Подсистема построения набора релевантных структур целевого языка на основе формальной семантической структуры обрабатываемого фрагмента текста, комплекса фразовых структур целевого языка (в общем случае – целевых языков) и словарей строит формальную семантическую структуру для построения выходного фрагмента текста. В связи с многовариантностью трансфера возможно построение набора выходных структур, характеризующихся как разной частотой использования в целевом языке, так и контекстом использования.

Подсистема построения выходного текста (текстов) на основе полученных выходных структур строит фрагмент выходного текста или его варианты. Также пользователю могут быть предложены подходящие примеры текста для построенных структур на основе существующих в системе коллекций «переводческой памяти». В зависимости от режима задаваемых целевых языков («многоязычности») и накопленных коллекций пользователю могут быть предоставлены разнообразные наборы текстов на разных языках.

Описания структур входного и целевых языков представляют собой независимые комплексы иерархически выстроенных правил фразовых структур соответствующих языков, выстроенные в единой нотации на основе когнитивной трансферной грамматики. Между собой комплексы фразовых структур разных языков связаны системой межязыковых отсылок, устанавливающих возможные направления трансфера от одного языка к другому. Подсистема фразовых структур языков обеспечивает ведение соответствующих архивов структур и вызов в подсистемы распознавания и трансфера необходимых пользователю комплексов структур языка.

Подсистема отображения, кроме обеспечения поэтапного вывода на экран результатов работы системы, предусматривает возможность коррекции пользователем построенных фразовых структур и выбор наиболее релевантной структуры на разных этапах работы.

На каждом этапе обработки метаописание модифицируется – структуры метаописания дополняется информацией, характеризующей текущий этап обработки, его состояние и возможные дальнейшие действия. После загрузки входного текста из файла или ввода его вручную в соответствующем окне ввода подсистема обработки входного текста сохраняет в исходном метаописании характеристики входного текста. Кроме того, отображаются выполненные над текстом операции, при необходимости дополняется сделанное ранее служебное описание текста описанием его фрагментации, а также сохраняется описание текущего уровня обработки введенного текста.

Подсистема распознавания структуры входного фрагмента дополняет метаописание характеристиками текущего этапа анализа и требуемых для анализа инструментов многоязычной лингвистической базы знаний. Там же сохраняются данные о полученных при анализе фрагмента текста результатах и о действиях пользователя над результирующими структурами. Аналогичным образом действуют все остальные подсистемы, получая в метаописании команды, запускающие их работу, и описания обрабатываемых данных, и в свою очередь модифицируя метаописание результатами своей работы и формируя очередные команды для продолжения работы над текущей задачей.

3 Метаданные управления и связи

Метаданные, используемые для управления работой многоязычной лингвистической базой знаний, образующие метаописание, можно разделить на следующие категории:

- метаданные состояния системы;
- метаданные процесса обработки;
- информационные метаданные.

Метаданные состояния системы описывают текущую конфигурацию многоязычной лингвистической базы знаний, работающего с системой пользователя, а также запрошенные им ресурсы. Описание текущей конфигурации включает в себя открытые пользователем окна и панели инструментария, их настройку и заданные объем и формат отображения исходных данных и результатов. Принадлежность пользователя к той или иной группе пользователей определяет объем его прав и возможностей при работе в системе, в первую очередь объем доступа к коллекциям базы знаний. В этой же секции метаописания хранится информация о текущих исходном и целевом языках, тематических областях, запрошенных словарях, коллекциях «переводческой памяти» и параллельных текстов. Одновременное использование нескольких словарей, а также различных коллекций, предполагает указание последовательности их применения (иерархии). Основные атрибуты описания состояния системы и их значения приведены в таблице 1. Атрибуты описания открытого окна приведены в таблице 2, характеристики открытых панелей в общем случае им аналогичны и характеризуют заданные параметры отображения панелей инструментов.

Табл. 1. Атрибуты описания состояния многоязычной базы знаний

Атрибут	Описание
Тип пользователя	Указывает, к какой группе пользователей принадлежит данный пользователь – администраторы, опытные или обычные пользователи, лингвисты и т.п.
Личные коллекции пользователя	Указывает место хранения личных коллекций и данных пользователя, своего рода «личное пространство» пользователя в системе

Число открытых окон	Указывает длину списка открытых (затребованных) пользователем окон
Список открытых окон	Содержит список открытых пользователем окон и их характеристики
Активное окно	Указывает последнее активное окно, с которым работал пользователь
Число открытых панелей	Указывает длину списка открытых (затребованных) пользователем панелей
Список открытых панелей	Содержит список открытых пользователем панелей и их характеристики
Активная панель	Указывает последнюю активную панель, с которой работал пользователь
Формат отображения текста	Указывает формат отображения исходного или целевого текста
Нотация фразовых структур	Указывает форму отображения распознанных фразовых структур для исходного фрагмента текста
Исходный язык	Указывает язык исходного текста (по умолчанию – русский)
Целевой язык	Указывает целевой язык, на который предполагается трансфер исходного текста. В режиме работы с фразовыми структурами может не использоваться
Научная область	Указывает терминологическую сферу языка научного документа или патента
Число открытых словарей	Указывает длину списка открытых (затребованных) пользователем словарей
Список открытых словарей	Содержит список открытых пользователем словарей и их характеристики
Порядок просмотра словарей	Указывает порядок просмотра словарей в процессе поиска требуемого слова

Табл. 2. Атрибуты описания открытого окна

Атрибут	Описание
Тип окна	Указывает, какого типа информация выводится в данное окно
Размещение	Указывает расположение окна на экране, заданное пользователем
Размер шрифта	Указывает размер шрифта, заданный пользователем

Метаданные процесса обработки текста хранят информацию о состоянии текущего этапа обработки и его фазе, результатах и параметрах дальнейшего процесса. Процесс обработки текста естественным образом разбивается на ряд этапов, описываемых метаданными:

- ввод исходного текста и его характеристик в пользовательскую среду,
- фрагментация текста на логико-структурные блоки (например, фразы),
- распознавание-разметку текущего фрагмента текста по морфологическим характеристикам составляющих его слов,
- построение его формальной семантической структуры или набора подходящих для данного фрагмента структур,
- поиск подходящих примеров в коллекциях базы знаний,
- построение набора фразовых структур целевого языка.

Особняком стоят метаданные процесса, сигнализирующие о состоянии паузы в работе программного комплекса базы в связи с ожиданием команды пользователя. Основными ожидаемыми командами могут быть ввод параметров работы системы, ввод текста на обработку и запуск процесса, указание предпочтительной фразовой структуры в построенном наборе и т.п. Результатом подачи той или иной команды является модификация выполняемой задачи и, соответственно, модификация текущего метаописания либо смена выполняемой задачи, сопровождаемая созданием нового метаописания.

Каждый этап обработки исходного текста состоит из ряда фаз. Например, разметка текущего фрагмента текста для последующего построения фразовой структуры фрагмента разбивается на ряд фаз:

- выделение слова,
- определение типа слова,
- поиск базовой словоформы по словарям и определение атрибутов, характеризующих текущее слово,
- сохранение списка возможных словоформ и соответствующих им наборов атрибутов для данного слова.

Ряд атрибутов описания состояния процесса обработки текста программным комплексом многоязычной лингвистической базы знаний приведен в таблице 3.

Табл. 3. Атрибуты текущего процесса обработки

Атрибут	Описание
Команда	Указывает введенную пользователем команду и заданные им параметры
Этап обработки	Указывает, на каком этапе обработки текста находится текущая задача
Фаза этапа	Указывает текущую фазу этапа обработки
Инструмент	Указывает вызванный на данных фазе или этапе блок или инструмент для выполнения задачи

Сформированный массив списков словоформ с атрибутами подается на вход подсистеме построения фразовых структур для дальнейшей многофазо-

вой обработки. Построение набора формальных семантических структур планируется выполнять на базе интегрированных в многоязычную лингвистическую базу знаний системы автоматического синтаксического анализа и модуля унификации, описанных в [8]. Набор построенных фразовых структур может быть предъявлен пользователю для принятия решения о дальнейшей работе программного комплекса (например, выбора варианта, наиболее релевантного исходному фрагменту текста).

В многоязычной лингвистической базе знаний используется метод моделирования языковых структур, который вырабатывался в процессе создания целого ряда проектов: интеллектуальных сред обработки знаний (ДИЕС, ИКС, ЛОГОС-Д) [5,6,11], а также экспериментальной системы машинного перевода Cognitive Translator [9]. Основной принцип построения логико-лингвистической модели - это декларативное представление правил синтаксического разбора на основе гибридной унификационно-порождающей грамматики. В системах ДИЕС, ИКС, ЛОГОС-Д значения языковых конструкций представлены на глубинно-семантическом уровне в виде расширенных семантических сетей (РСС) и продукционных правил. При этом используется подход «синтаксической нормализации» [5,6]. Суть этого подхода состоит в том, что структуры с неличными формами глаголов приводятся к форме сложноподчиненных предложений с придаточными условия и причины. Это позволяет отобразить их в глубинно-семантические пропозициональные структуры.

В системе машинного перевода Cognitive Translator был разработан специализированный внутренний язык представления лингвистических знаний, поддерживающий механизм отображения, и отношений зависимости, и составляющих [9]. Лингвистическая база знаний ИНТЕРТЕКСТ предназначена для поддержки как поверхностно-синтаксических отображений, так и глубинных представлений языковых структур. Например, синтаксические правила для языковых конструкций со значением причины и условия представлены следующим образом [4]:

SYNRULE S/if_cl: S{fin+, sub_conj+} -> S{fin+,sub_conj+} PUNCT+ S{fin+,cl+}

Since there is a continual loss of micrometeoritic space because of the radiation effects, there must be a continual replenishment: otherwise micrometeorites would have disappeared from interplanetary space.

Поскольку имеет место постоянное сокращение микрометеоритного пространства по причине радиационных эффектов, должно быть постоянное пополнение: иначе микрометеориты исчезли бы из межпланетного пространства. (В связи с тем, что постоянно происходит сокращение микрометеоритного пространства, вызванное радиационными эффектами, по всей вероятности, происходит его постоянное пополнение, в противном случае, микрометеориты уже давно исчезли бы из межпланетного пространства).

SYNRULE S/vppt_if: S{fin-,sub_conj-} -> S{fin-,sub_conj-} PUNCT+ S{fin+,cl+}

Debugged, the program ran

Исправленная, программа заработала (Поскольку она была исправлена, программа заработала).

Информационные метаданные служат для передачи информационных сигналов от одной подсистемы или функционального блока другим подсистемам и функциональным блокам программного комплекса базы знаний, а также для организации информационных сообщений пользователю. Информационные метаданные хранят информацию о текущем состоянии процесса, обнаруженных конфликтных ситуациях (ошибках) и их разрешении, передаче задачи следующей подсистеме или возврат назад для повторной обработки, а также сообщения внутри подсистемы от одного функционального блока другому. Информационные метаданные могут обладать сложной структурой и включать в себя ряд параметров, характеризующих возникшую ситуацию и точку ее генерации. Основные атрибуты информационных сигналов системы приведены в таблице 4.

Табл. 4 Атрибуты информационных сигналов процесса обработки

Атрибут	Описание
Завершение этапа	Указывает текущее состояние выполняемого/завершенного этапа
Завершение фазы	Указывает текущее состояние выполняемой/завершенной фазы
Следующий инструмент	Указывает требуемый далее инструмент
Конфликт	Указывает параметры ситуации, неразрешенной системой

4 Заключение

Применяемые в многоязычной лингвистической базе знаний метаданные, кроме обычного описания данных и коллекций, хранят текущее описание состояния базы знаний и выполняемых в ней действий, что позволяет организовать связи между подсистемами программного комплекса в зависимости от выполняемых заданий и управление их функционированием.

Наборы метаданных, описывающие текущее состояние многоязычной лингвистической базы знаний в процессе выполнения задания, организуются в виде метаописания, формирующегося в начале сеанса работы. В процессе работы базы метаописание модифицируется, чтобы отображать текущее состояние системы, что позволяет при необходимости приостановить заданный процесс и впоследствии продолжить его с того же самого места.

Каждая функциональная подсистема или блок базы знаний, получая доступ к метаописанию, считывает из него информацию о текущем состоянии работы над указанными данными и своих действи-

ях. В свою очередь, подсистема в процессе своего функционирования модифицирует метаописание, занося в него результаты выполненного этапа или его фазы и входные параметры для дальнейших действий программного комплекса базы, которые служат командами для других подсистем. За счет организации через метаописание управления многоязычной лингвистической базой знаний достигается гибкая самонастройка всего комплекса на особенности каждой решаемой задачи.

Литература

[1] Козеренко Е.Б. Лингвистические аспекты информатики // Системы и средства информатики. Специальный выпуск Научно-методологические проблемы информатики. Москва, ИПИРАН, 2006, 88-111.

[2] Козеренко Е.Б. Логико-статистические методы представления языковых структур в машинном переводе. // Труды Международной конференции Диалог'2005 "Компьютерная лингвистика и интеллектуальные технологии", М.: "Наука", 2005.

[3] Козеренко Е.Б. Моделирование переноса функциональных значений для англо-русского машинного перевода. // Труды Международной конференции Диалог'2004 "Компьютерная лингвистика и интеллектуальные технологии", М.: "Наука", 2004.

[4] Козеренко Е.Б. Моделирование языковых структур со значением условия и причины для англо-русского лингвистического процессора. // Труды Международной конференции Диалог'2003 "Компьютерная лингвистика и интеллектуальные технологии", М.: "Наука", 2003, стр. 312-316.

[5] Козеренко Е.Б. Унифицированные категориально-функциональные представления для синтаксической разметки полнотекстового документа//Системы и средства информатики. М.: Наука, 2002 Вып.11.

[6] Козеренко Е.Б. Функциональная семантика в компьютерных решениях//Труды Международного семинара «Диалог'2002» по компьютерной лингвистике и интеллектуальным технологиям. г. Протвино Московской обл. 2002 т.1. Стр. 218-226.

[7] Лунева Н.В. Архитектура и метаданные многоязычной лингвистической базы знаний // Системы и средства информатики. Вып.17. Москва, ИПРАН, 2007, стр. 317-336.

[8] Перекрестенко А.А. Разработка и программная реализация системы автоматического выделения синтаксических групп для естественных языков знаний // Системы и средства информатики. Вып.17. Москва, ИПРАН, 2007, стр. 273-291.

[9] Kozerenko E.B. Cognitive Approach to Language Structure Segmentation for Machine Translation Algorithms // Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications, June, 23-26, 2003, Las Vegas, USA.// CSREA Press, 2003. P. 49-55.

[10] Kozerenko E.B. INTERTEXT: A Multilingual Knowledge Base for Machine Translation // Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications, June, 25-28, 2007, Las Vegas, USA.// CSREA Press, pp. 238 - 243, 2007.

[11] Kozerenko E.B. Portable Language Egineering Solutions for Multilingual Processors // Proceedings of the International Conference on Artificial Intelligence IC-AI'02// CSREA Press, 2002, pp. 447-453.

Metadata in the management system of multilingual linguistic knowledge base

N.V. Luneva

The use of metadata in the management system of multilingual linguistic knowledge base, their role in the interaction of functional subsystems and units of software of knowledge base are considered. Metadata are used to describe external and internal data of the system, their current status as well as the status of the system and to exchange information between functional subsystems and units of multilingual linguistic knowledge base.