ИНТЕГРАЛЬНАЯ ТЕХНОЛОГИЯ РАЗРЕШЕНИЯ ОМОНИМИИ В СИСТЕМЕ АНАЛИЗА ТЕКСТОВЫХ ДОКУМЕНТОВ "ЛОТА"[1]

INTEGRAL TECHNOLOGY OF HOMONYMY DISAMBIGUATION IN THE LOTA TEXT MINING SYSTEM

Невзорова О.А. *(olga.nevzorova@ksu.ru), НИИММ им. Н.Г. Чеботарева, Казань*
Невзоров В.Н. *(nevzorov@mi.ru), Казанский государственный технический университет им. А.Н. Туполева*
Зинькина Ю.В. *(zjuliv@mail.ru), Казанский государственный университет*
Пяткин Н.В. *(nikolaip@mail.ru), НИИММ им. Н.Г. Чеботарева*

В статье описывается интегральная технология разрешения многозначности, реализованная в системе анализа текстовых документов "ЛоТА". Технология содержит совокупность методов разрешения омонимии и схему их взаимодействия.

1. Введение

Специализированная система обработки текстовых документов "ЛоТА" [1] является системой класса Text Mining. Система предназначена для анализа специализированных текстов "Логика работы", описывающих логику работы сложной технической системы в различных режимах функционирования. Основной задачей анализа является извлечение из данных текстов информационной модели алгоритмов, решающих определенную задачу в определенной проблемной ситуации, и контроль структурной и информационной целостности выделенной схемы алгоритмов.

Информационная модель алгоритма включает:

- описание входного информационного потока (типы информационных сигналов или семантическое описание информационного потока с указанием источника информации - конкретный алгоритм, конкретное измерительное устройство);
- описание процессов преобразования входных данных в выходные (допустимый способ разрешения проблемы);
- описание выходного информационного потока (типы информационных сигналов или семантическое описание информационного потока с указанием точки приема информации).

Решение основной задачи обеспечивается комплексом технологий обработки текстов, включающих:

- технологии морфосинтаксического анализа;
- технологии семантико-синтаксического анализа;
- технологии взаимодействия с прикладной онтологией.

Указанная сумма технологий формируется на основе центрального ядра – прикладной онтологии (в дальнейшем, авиаонтология), обеспечивающей согласованное взаимодействие различных программных модулей. Авиаонтология концептуально описывает предметную область информационного обеспечения различных полетных режимов антропоцентрических систем [2]. Авиантология представляет собой сеть понятий предметной области. Текущий размер онтологии - свыше 1600 понятий (около 5000 текстовых входов понятий). Авиаонтология относится к классу лингвистических (лексических) онтологий и предназначена для встраивания в различные лингвистические приложения.

Программный комплекс состоит из трех взаимодействующих подсистем: подсистемы лингвистического анализа технических текстов "Анализатор", подсистемы ведения онтологии "OntoEditor+" и подсистемы "Интегратор". Взаимодействие подсистем реализовано на базе технологии "клиент-сервер", причем в различных подзадачах подсистемы выступают в различных режимах (режим сервера или режим клиента).

Инструментальная система визуального проектирования "OntoEditor+" [4] является специализированной СУБД. Система предназначена для ручного редактирования онтологий, хранящихся в реляционной базе данных в формате TPS, а также обслуживания запросов пользователей и внешних программ. Новые возможности системы обеспечиваются функциональным набором "Лингвистический инструментарий", посредством которого реализуется встраивание прикладной онтологии в лингвистические приложения. Наиболее типичными задачами, решаемыми с помощью инструментария системы "OntoEditor+", являются: изучение структурных свойств при-

[1] Работа выполнена при поддержке Российского Фонда Фундаментальных Исследований, грант № 05-07-90257.

кладной онтологии с помощью исследовательского инструментария системы “OntoEditor+”; построение лингвистической оболочки прикладной онтологии; задача покрытия текста онтологическими входами; построение выводов по прикладной онтологии и др.

Подсистема “Анализатор” реализует основные этапы лингвистической обработки текста (графематический, морфосинтаксический и частичный синтаксический анализ). В статье будет рассмотрена интегральная технология разрешения многозначности, которая ориентирована прежде всего на разрешение функциональной, морфологической и лексической омонимии.

Подсистема “Интегратор” исполняет внешний запрос на извлечение знаний из текста. Структура внешнего запроса содержит компоненты информационной модели алгоритма. Внешний запрос интерпретируется при взаимодействии с подсистемой “OntoEditor+” как структура, привязанная к прикладной онтологии. Выделение компонент информационной модели происходит на основе механизмов отождествления элементов дерева сегментов входного текста (взаимодействие с подсистемой “Анализатор”) и элементов структуры запроса (взаимодействие с подсистемой “OntoEditor+”).

2. Интегральная технология разрешения многозначности

2.1. Предыдущие работы

Разрешение функциональной, морфологической и лексической омонимии является одной из актуальных задач обработки текстов. К настоящему времени сформирована основная парадигма методов снятия омонимии, которая включает методы, основанные на правилах; методы машинного обучения, использующие вероятностные модели; гибридные методы. В мировой литературе представлены большое число публикаций по разрешению омонимии на основе статистического подхода для основных европейских языков. Эта технология последние годы активно развивается для русского языка [5-7], прежде всего благодаря проекту “Национальный корпус русского языка”, который предоставляет размеченный подкорпус русского языка для настройки алгоритмов машинного обучения. Подход, основанный на правилах, является чрезвычайно трудоемким, требует проведения тщательной лингвистической экспертизы каждого типа омонимии. Несмотря на большой исторический возраст, этот метод для русского языка в полной мере не описан в открытой литературе, некоторые принципиальные идеи и реализация представлены в [9]. В [6] даны сравнительные оценки различных модулей разрешения омонимии, построенных на основе статистических методов и метода, основанного на правилах. В целом, оценки для случая полного разрешения функциональной омонимии достаточно близки 97,26 % и 96, 87 %. Можно предположить, что неразрешенные примерно 3 % относятся к синтаксически сложным случаям и многие авторы сходятся во мнении, что наиболее эффективным является использование гибридных технологий разрешения омонимии. Однако следует отметить, что полученные оценки даны для классификации типов омонимии, принятой в Национальном корпусе русского языка. Эта классификация в ряде конкретных случаев функциональной омонимии расходится с другими классификациями. Примеры ряда таких расхождений можно найти в [3]. В качестве примера можно привести грамматические характеристики омонима *раз* (в НКРЯ – наречие/союз/существительное; уточненный тип - наречие/союз/существительное/числительное/частица).

2.2. Структура технических текстов “Логика…“

Тексты “Логика” являются реальными техническими текстами, имеющими сложную синтаксическую структуру. Тексты содержат большое количество аббревиатур, в том числе авторских, предложения с перечислениями, омонимию различных типов и т.п. Этап лингвистического анализа поддерживается рядом стандартных лингвистических ресурсов, среди которых можно выделить грамматический словарь, построенный на основе грамматического словаря Зализняка А.А. существенно дополненный наречиями, специальной лексикой; словарь аббревиатур и стандартных сокращений; словари устойчивых словосочетаний и др. К нестандартным ресурсам, поддерживающих лингвистический анализ текстов относятся лингвистическая оболочка авиаонтологии и индексируемая база устойчивых коллокаций предметной области. Фактически, эти ресурсы обеспечивают основные технологии разрешения многозначности в системе, а именно разрешение многозначности на основе индексируемой базы устойчивых коллокаций, а также разрешение многозначности на основе лингвистической оболочки авиаонтологии.

В течение последних лет авторский коллектив разрабатывает универсальную технологию разрешения функциональной многозначности омонимов на основе метода контекстных правил [3]. Данная технология основывается на тщательных лингвистических исследованиях синтаксического поведения омонимов, уточнения их грамматических характеристик. Выбор метода был обусловлен требованиями задачи (необходимость полного снятия всех типов омонимии), а также специфическим характером текстов заданной проблемной области.

Данная прикладная задача позволила выявить ряд актуальных проблем лексикографического описания функциональных омонимов русского языка. Были предложены новые статистические основания для классификации омонимов и выделения подтипов внутри типов омонимии, а также начаты работы по разработке нового словаря функциональных омонимов русского языка на основе корпусных исследований. Метод контекстного разрешения омонимии является базовым методом в интегральной технологии разрешения омонимии в системе "ЛоТА". Однако, практические задачи системы выявили ряд важных аспектов лингвистического анализа, которые стимулировали развитие новых методов разрешения многозначности. В первую очередь, были выполнены работы, связанные с получением количественных и качественных оценок специализированной текстовой базы документов системы. Анализ позволил выявить количественные оценки и распределение омонимов по типам. Другая важная оценка была получена при анализе типовых контекстов омонимов. Проведенные исследования выявили степень омонимичности технических текстов (в среднем, 15-20 % омонимов); были построены частотные списки омонимов, а также типовые контексты частотных омонимов. Эти результаты легли в основу новых прикладных технологий разрешения многозначности.

Таким образом, интегральная технология разрешения многозначности, разрабатываемая в системе "ЛоТА" включает следующие методы:

- метод контекстного разрешения функциональной омонимии;
- метод разрешения функциональной, грамматической и лексической омонимии на основе индексируемой базы устойчивых коллокаций;
- метод разрешения функциональной, грамматической и лексической омонимии на основе лингвистической оболочки онтологии.

2.3. Метод контекстного разрешения функциональной омонимии

Метод контекстного разрешения функциональной омонимии сводится к разработке для каждого функционального типа омонимии группы правил, задающих синтаксический контекст разрешения омонима, и построение управляющей структуры группы, определяющей порядок применения правил. В работе [3] были подробно описаны основные достоинства и недостатки данного метода, приведены конкретные структуры обобщенных правил для разрешения функциональной омонимии некоторых типов. Метод контекстного разрешения применяется на этапе морфосинтаксического анализа текста, так как достаточно часто при разрешении омонимии используется синтаксический метод построения однородных групп. Разрешение омонима за границами однородной группы позволяет учитывать не только локальный контекст омонима, но и дальнее окружение. Это обстоятельство, несомненно, является одним из главных преимуществ метода. Как отмечалось в [3] текущая оценка точности метода составляет 95 %. В интегральной технологии метод контекстного разрешения является базовым и применяется последним в группе методов разрешения многозначности.

2.4. Метод разрешения многозначности на основе индексируемой базы устойчивых коллокаций

Для реализации метода разработана интегрированная программная технология построения индекса базы контекстов омонимов. К числу наиболее идейно близких результатов относятся результаты, изложенные в [8]. Рассматриваемый метод позволяет максимально учитывать специфические особенности тематики текстов, прежде всего лексические и терминологические. Разработанная программная технология включает модули создания и ведения индекса омонимов, модуль согласования индексной базы с основным лингвистическим ресурсом – грамматическим словарем, и механизмы выполнения внешних запросов по разрешению (поиску) типовых омонимических контекстов во входном тексте на основе индекса омонимов. Разработанная технология реализована на основе взаимодействия основных подсистем: подсистемы "OntoEditor +" и подсистемы "Анализатор".

Для эффективного встраивания в лингвистические приложения система "OntoEditor+" поддерживает группу протоколов информационного обмена с внешними программными модулями системы и внешними словарными базами данных, обеспечивая работу в режиме клиент-сервер. Разрешение многозначности (функциональной, морфологической и лексической) во входных текстах происходит на основе механизма распознавания контекстов омонимов, зафиксированных в индексируемой базе контекстов.

Разработаны три основных механизма пополнения индексируемой базы контекстов функциональных омонимов:

- ручной ввод и редактирование данных по типовым контекстам омонимов;
- импорт типовых контекстов омонимов из текстового файла, подготовленного в специальном формате представления данных;

- импорт типовых контекстов омонимов, обнаруженных специальными механизмами поиска подсистемы “Анализатор”.

Данный механизм организован как запрос к подсистеме “Анализатор“ с передачей ему от подсистемы “OntoEditor+” текстового корпуса, по которому проводится поиск. В процессе обработки подсистема “Анализатор” передает подсистеме “ OntoEditor+” информацию об обнаруженных контекстах омонимов, которая записывается либо в индекс омонимов либо в автоматическом режиме, либо в режиме диалога с оператором. Отличительной особенностью режима диалога является режим самообучения, который реализуется с использованием механизма журнала событий. В данном журнале в зависимости от его настройки фиксируются те или иные важные события в системе, например, изменение информации в индексе омонимов или операции взаимодействия с подсистемой “Анализатор”. В режиме самообучения сохраняется и контролируется последовательность ранее сгенерированных диалогов, что обеспечивает генерацию только уникальных диалогов на разрешение омонимии без повторений.

На основе экспериментальной текстовой коллекции, а также ряда лингвистических ресурсов, среди которых наиболее существенными являются Национальный корпус русского языка (www.ruscorpora.ru), а также Русский ассоциативный словарь в 2-х томах (Караулов Ю.Н., Черкасова Н.В. и др. - М.: ООО “Изд.-во Астрель”, 2002**)** построена база разрешающих коллокаций функциональных омонимов (текущий размер базы составляет около 30000 коллокаций). Разработан программный модуль, обеспечивающий генерацию экземпляров разрешающих коллокаций по их моделям при построении индекса базы функциональных омонимов. Модель коллокации задает разрешающий контекст лексического или функционального омонима. В настоящее время по указанным лингвистическим ресурсам построено около 2000 моделей коллокаций. Модель коллокаций состоит из двух частей. В первой части представлены словоформы - компоненты словосочетания (как правило, бинарные или тернарные), во второй части содержатся кодовые параметры внутреннего описания словоформ по грамматическому словарю, а также позиция и расстояние разрешающей словоформы относительно омонима. Такая модель позволяет реализовать генерацию всех разрешающих контекстов, которые различаются формой разрешающей словоформы. Например, модель коллокации *‘относительно короткий’* (разрешение функционального омонима *относительно* как наречие) расширяется всей парадигмой прилагательного ‘*короткий*‘. Статистический анализ типов моделей коллокаций позволил выявить наиболее частотные типы, к которым относятся следующие модели:

- омоним (наречие/краткое прилагательное) + глагол (разрешающая словоформа), например, ‘*эффективно действовать’* ;

- омоним (существительное/прилагательное) + существительное, например, ‘*ближний бой*‘ .

Метод разрешения омонимии на основе моделей коллокаций эффективно применяется при разрешении сложных случаев омонимии, например для омонимов *это, все*/*всё* выделены более 200 разрешающих коллокаций.

2.5. Метод разрешения многозначности на основе лингвистической оболочки онтологии.

Лингвистический инструментарий подсистемы “OntoEditor+” обеспечивает встраивание онтологии в различные приложения, связанные с обработкой текстов. Лингвистический инструментарий реализует функции загрузки корпуса текстов; автоматическое ведение статистики по различным объектам корпуса; функции предсинтаксической обработки текста (сегментация предложений, распознавание аббревиатур, разрешение омонимии на основе специальных протоколов взаимодействия с внешними словарными ресурсами); построение лингвистической оболочки онтологии; распознавание терминов прикладной онтологии во входном тексте (задача покрытия). Сопряжение онтологического и лингвистического (грамматического) ресурсов реализуется через механизмы лингвистической оболочки онтологии. Лингвистическая оболочка онтологии создается с помощью разработанного программного инструментария, посредством которого фиксируется грамматическая информация об онтологических концептах и их текстовых формах. Каждый онтологический вход (как правило, многословный термин) снабжается соответствующей грамматической информацией, при этом для омонима разрешается соответствующая (функциональная, лексическая, морфологическая) омонимия. Грамматическая информация передается в подсистему “OntoEditor+” от подсистемы “Анализатор” на основе специальных протоколов взаимодействия. Разрешение лексической, функциональной и морфологической омонимии выполняется на основе специальных диалогов с экспертом-лингвистом. Отдельные процедуры реализуют проверки словоформ в составе терминологического входа на согласованность их грамматических характеристик, также осуществляется контроль достоверности словарной информации. Контроль достоверности обеспечивает отслеживание изменений, как в составе грамматического словаря, так и в составе онтологии. Учитывая сложность и многоступенчатость вышеперечисленных процедур, в подсистеме “Ontoeditor+” разработан мастер построения лингвистиче-

ской оболочки, который вызывается командой основного меню.

Механизм разрешения омонимии на основе лингвистической оболочки онтологии связан с решением задачи распознавания в тексте онтологических входов (задачи покрытия текста). Для каждого распознанного онтологического входа, содержащего омоним, передается информация о грамматических характеристиках омонима в контексте онтологического входа. Так, например, при распознавании в тексте онтологического концепта *'ближний бой',* лингвистическая оболочка данного концепта содержит информацию о грамматических характеристиках каждого омонима в составе данного входа. Метод позволяет разрешать функциональную, морфологическую, а также лексическую омонимию.

2.6. Взаимодействие методов разрешения омонимии

Интегральная технология разрешения омонимии включает три вышеуказанных метода разрешения омонимии. Взаимодействие методов в решении задачи разрешения омонимии обеспечивается через взаимодействие основных подсистем системы "ЛоТА".

Подсистема "OntoEditor+" обеспечивает реализацию метода разрешения многозначности на основе устойчивых коллокаций и метода разрешения многозначности на основе лингвистической оболочки онтологии. При разработке этих методов используется инженерный подход, позволяющий выделять типовые частотные языковые ситуации, которые активно используются в техническом языке. По сути, при разрешении многозначности на основе этих методов используются общие и специальные знания системы, хранящиеся в различных базах данных.

Подсистема "Анализатор" обеспечивает реализацию метода разрешения омонимии на основе контекстных правил, т.е. фактически используются лингвистические знания системы. Этот метод является универсальным, не зависит от специфики предметной области и обеспечивает в текущей версии точность распознавания не ниже 95 %. Однако, для данного метода существуют крайне сложные типы функциональной омонимии, например, тип "частица/союз". Разрешение данной омонимии возможно во многих случаях лишь после завершения полного синтаксического анализа.

Взаимодействие подсистемы "OntoEditor+" и подсистемы "Анализатор" осуществляется на основе специальных протоколов взаимодействия. При применении интегральной технологии разрешение многозначности происходит в два этапа. На первом этапе подсистема "Анализатор" (клиент) передает запрос на разрешение омонимии входного текста подсистеме "OntoEditor+" (сервер). Подсистема "OntoEditor+" возвращает подсистеме "Анализатор" информацию о разрешенных омонимах на основе своих методов. На втором этапе подсистема "Анализатор" разрешает омонимию оставшихся неразрешенных омонимов на основе метода контекстных правил.

3. Заключение

Интегральная технология разрешения многозначности эффективно применяется на этапе предсинтаксического анализа в системе "ЛоТА". По существу, интегральная технология представляет собой сочетание инженерного и лингвистического подхода к решению поставленной задачи. В основе проектирования интегральной технологии лежат процессы скоординированного взаимодействия различных языковых уровней, прежде всего онтологического уровня (обеспечивающего системные модели знаний о мире) и различных языковых уровней (морфологического и синтаксического). В системе реализован эффективный механизм взаимодействия различных подсистем, обеспечивающих реализацию различных методов в составе интегральной технологии.

Список литературы

1. Невзорова О.А., Федунов Б.Е. Система анализа технических текстов "ЛоТА": основные концепции и проектные решения. // Изв. РАН. Теория и системы управления.– 2001. № 3. С. 138-149.
2. Добров Б.В., Лукашевич Н.В., Невзорова О.А., Федунов Б.Е. Методы и средства автоматизированного проектирования прикладной онтологии // Известия РАН. Теория и системы управления. М.: 2004. № 2. С. 58-68.
3. Невзорова О.А., Зинькина Н.В., Пяткин Н.В. Метод контекстного разрешения функциональной омонимии: анализ применимости // Труды межд. конф. Диалог'2006. М., Наука, 2006. С. 399 – 402.
4. Невзорова О.А., Невзоров В.Н. Система визуального проектирования онтологий "OntoEditor": функциональные возможности и применение //IX национальная конференция по искусственному интеллекту с международным участием КИИ-2004. М.: Физматлит, 2004. Т. 3. С.937-945.

5. Зеленков Ю.Г., Сегалович И.В., Титов В.А., Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов // Компьютерная лингвистика и интеллектуальные технологии. Труды международного семинара Диалог'2005., 2005. С. 188-197.
6. Сокирко А.В., Толдова С.Ю. Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка (скрытая модель Маркова и синтаксический анализатор именных групп) // Интернет-математика-2005. http://company.yandex.ru/grant/list.xml
7. Jiri Hana and Anna Feldman, Portable Language Technology: The case of Czech and Russian. In Proceedings from the Midwest Computational Linguistics Colloquium, June 25-26, 2004, Bloomington, Indiana.
8. Кобрицов Б.П., Ляшевская О.Н., Шеманаева О.Ю. Снятие лексико-семантической омонимии в новостных и газетно-журнальных текстах: поверхностные фильтры и статистическая оценка. // Интернет-математика-2005. http://company.yandex.ru/grant/list.xml
9. Кобзарева Т.Ю., Афанасьев Р.Н. Универсальный модуль предсинтаксического анализа омонимии частей речи в РЯ на основе словаря диагностических ситуаций // Труды междунар. конференции Диалог'2002. М., 2002. С. 258-268.