OSTIS-2013
(Open Semantic Technologies for Intelligent Systems)

УДК 004.912

ПРОГРАММНЫЙ ИНСТРУМЕНТАРИЙ РАЗРАБОТКИ ЛИНГВИСТИЧЕСКИХ РЕСУРСОВ

Загорулько М.Ю., Сидорова Е.А.

Институт систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук, г. Новосибирск, Россия

zagorulko_maxim@yahoo.com

lena@iis.nsk.su

В работе рассматривается комплекс программных средств, предназначенных для проведения лингвистических исследований и разработки лингвистических ресурсов: корпусов текстов и предметно-ориентированных словарей. Описаны словарные системы, предназначенные для создания терминологических, лексико-семантических и семантико-синтаксических словарей.

Ключевые слова: лингвистический ресурс; терминологический словарь; корпус текстов; конкорданс.

Введение

Разработка естественно-языковых интерфейсов для обеспечения пользователей информационных систем более удобными средствами доступа, поиска и представления информации требует решения разнообразных лингвистических задач. Сложность задачи зависит от лингвистических потребностей разработчиков информационной системы. Это может быть задача извлечения терминологии и построения терминологического словаря или задача обработки текста на основе лексикографического (словарного) ресурса или реализация полноценного семантического анализа текста и автоматического извлечения информации. Решение любой из перечисленных задач требует наличия определенных лингвистических ресурсов, создание которых (в случае их отсутствия) является отдельной задачей.

Совершенно необходимым ресурсом для решения практически любой задачи является морфологический словарь одного или нескольких языков. Остальные лингвистические ресурсы, как правило, предметно-ориентированы на ту сферу деятельности, в рамках которой создается информационная система. Можно выделить следующие типы лексикографических ресурсов.

- Терминологические словари, в состав которых входят как однословные, так и многословные термины [Большаков, 2002];
- Тезаурусы, содержащие знания о языке в проекции на конкретную сферу деятельности [Лукашевич, 2011];
- Лексико-семантические и семантико-синтаксические словари, основанные на шаблонных описаниях (например, словари моделей управления [Волкова и др., 1998]) и др.

Неотъемлемой частью процесса создания лингвистических ресурсов и разработки методов решения поставленных лингвистических задач, является проведение предварительного лингвистического исследования тех материалов, с которыми в дальнейшем будет работать система. Лингвистические исследования с помощью современных компьютерных технологий проводятся корпусными методами [Захаров и др., 2011] и их часто называют «объективными» или квантитативными. Результатом такого исследования может стать аннотированный корпус текстов, который в дальнейшем служит для автоматизированного создания лексикографических ресурсов. Лексикографическое направление является на данный момент основным в корпусной лингвистике, это связано с тем, что основным инструментом исследования корпуса, является построение конкордансов (т.е. совокупности контекстов слов), которые изначально, еще в «доцифровую» эпоху были основой для создания словарей [Богданова, 2010].

В работе рассматривается комплекс программных средств, предназначенных для проведения лингвистических исследований и разработки лингвистических ресурсов: корпусов

текстов и предметно-ориентированных словарей. Вопросы применения данных лингвистических ресурсов для автоматического анализа текста и, связанные с этим лингвистические модели, останутся за рамками этого доклада.

1. Лингвистические потребности информационных систем

В процессе разработки естественно-языковых модулей возникают различные лингвистические задачи, решение которых требует применения разнообразных лингвистических ресурсов.

1.1. Лингвистические задачи

Рассматривая общую задачу автоматической обработки текста можно выделить несколько подзадач, решение которых связано с созданием и использованием лингвистических ресурсов.

(1) Как уже было отмечено, проведение *лингвистических исследований текстов* необходимо как для выработки основных принципов последующей разработки лингвистических ресурсов, необходимых для решения поставленных задач, так и конкретного наполнения этих ресурсов.

(2) Решение задачи *извлечения терминологии* служит не только для дальнейшей поддержки словарного анализа текста, но и для предварительного экспертного исследования предметной области (ПО), выявления основных классов понятий и логических взаимосвязей между понятиями, которые соответствуют введенным терминам, создания модели ПО, формирования структуры БД системы.

(3) Одним из главных назначений компьютерных лингвистических ресурсов является поддержка одного из этапов *автоматической обработки текста*, целью которого является извлечение из текстов словарных концептов.

(4) В более общем виде задача обработки текста подразумевает *извлечение информации*. Для решения этой задачи используют дополнительные модели и знания о согласованиях и взаимосвязях языковых единиц и элементов ПО с учетом жанровых особенностей документов.

Для решения указанных лингвистических задач разрабатываются лингвистические ресурсы, а также методы и средства их автоматизированного создания и поддержки.

1.2. Лингвистические ресурсы

В нашей технологической цепочке мы рассматриваем следующие типы лингвистических ресурсов.

(1) Корпус текстов – подборка текстов определенного жанра, тематика которых соответствует заданной ПО. Корпус может содержать лингвистическую разметку, представляющую собой информацию, полученную автоматически при анализе текстов, либо приписанную экспертом вручную. Основное назначение корпуса – автоматизация создания других лингвистических ресурсов.

(2) Универсальные и предметные словари, содержащие перечень минимальных единиц языка, терминов и устойчивых терминологических словосочетаний, используемых при описании значимой для разработчиков системы информации, а также жанровую лексику, описываемую лексическими шаблонами, для извлечения нестандартно представленной в тексте лексики. В рамках словарей определяется набор специфичных для данного подъязыка лингвистических знаний: морфологические классы (определение лексемы для найденной в тексте словоформы), правила формирования многословных терминов (синтаксические шаблоны).

(3) Семантический словарь, формирующий семантические признаки и отношения на лексиконе. Семантический словарь включает целевые тезаурусы (например, справочно-информационный тезаурус, тезаурусы для анализа текста, для поддержки информационного поиска, для перевода и т.п.), а также семантико-синтаксические словари, которые ограничивают синтаксическую сочетаемость и проверяет согласованность грамматических и семантических признаков терминов (вершин синтаксических групп) в соответствии с правилами согласования и управления. Эти знания могут быть заданы с разной степенью подробности в зависимости от требований и возможностей разработчиков системы.

(4) Набор описаний жанровых структур текста в совокупности с логическим представлением текста образуют модели документов, соотнесенных с тем или иным типом или жанром текстовых ресурсов.

(5) Знания о согласовании имеющихся лингвистических знаний с предметными знаниями. С этой целью термины группируются в семантические группы, которые в свою очередь также согласуются с элементами онтологии либо непосредственно, либо в соответствии с определенной схемой (схемой факта).

Для решения поставленных разработчиками задач и создания необходимых для их решения ресурсов нами разработан набор программных средств.

Данные инструменты поддерживают как создание лингвистических ресурсов, так и дальнейшее их использование в задачах автоматической обработки текста.

1.3. Схема взаимодействия модулей программного инструментария

Вся технологическая цепочка объединяет три функциональных компонента, отвечающих за управление корпусом (корпус-менеджер), построение словарей, а также за извлечение

информации (рисунок 1). Каждый из компонентов состоит из отдельных программных модулей, имеющих пользовательский интерфейс и определенный формат обмена данными между другими модулями.

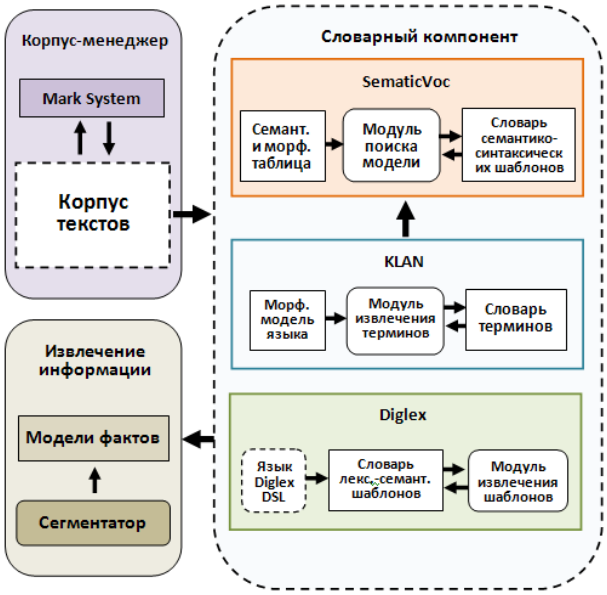

Рисунок 1 – Компонентная архитектура инструментария

Рассматривая весь процесс разработки лингвистических ресурсов и решения поставленных лингвистических задач, можно выделить следующие этапы.

(1) *Обработка корпуса текстов.* Помимо развитых средств управления коллекцией текстовых документов корпус-менеджер предоставляет возможность многоуровневой разметки текста, на основе которой реализуется автоматизированное наполнение предметных словарей.

(2) *Построение словарей*. Словарные модули поддерживают разработку словарей, предоставляют необходимые редакторы, реализуют, где это возможно, методы автоматического начального наполнения словарей на основе корпуса текстов, поддерживают дальнейшее использование словарей для обработки текста.

(3) *Извлечение информации* [Загорулько и др., 2009] предполагает наличие дополнительных моделей (модели документов, модели фактов, онтологии ПО) и специализированных модулей (в данной работе не рассматриваются).

2. Инструменты создания словарей

Наш инструментарий поддерживает создание нескольких типов словарей.

- Терминологические словари (система KLAN), содержащие предметную и необходимую универсальную лексику.
- Лексико-семантические словари (система DigLex), содержащие как нетерминологические единицы, имеющие регулярную структуру (например, номер телефона, дата, инициалы) так и специфические термины предметной области, отсутствующие в базовом словаре и/или имеющие сложную структуру (например, полные названия организаций, событий).
- Семантико-синтаксические словари (система SemanticVoc), содержащие модели согласования терминов или классов терминов.

Все словарные системы снабжены пользовательским редактором ресурса, модулем тестирования, осуществляющим текстовый анализ выбранного текста и визуализацию результатов, а также программным API, позволяющим использовать разработанный лингвистом-экспертом ресурс автономно.

2.1. Система KLAN

Система KLAN предназначена для создания предметных словарей (рисунок 2) и позволяет включать в терминологический лексикон грамматическую и статистическую информацию, используемую как для наполнения словаря на основе обучающей выборки, так и для автоматического извлечения словарных терминов из текста.

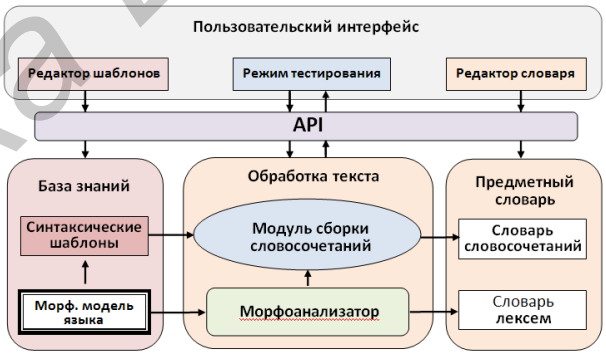

Рисунок 2 – Архитектура системы KLAN

В качестве термина может выступать слово или словосочетание, в общем случае, произвольной длины и состава [Лукашевич и др., 2008]. Для поиска новых словосочетаний используются синтаксические шаблоны, описываемые в терминах морфологических классов и характеристик [Большакова и др., 2010]. Синтаксический шаблон задает иерархическое представление синтаксической структуры словосочетания, где связь между элементами словосочетания осуществляется на основе морфологических характеристик. Система имеет пользовательский интерфейс для редактора шаблонов, позволяющего задать любую синтаксическую конструкцию.

Система поддерживает экспертную настройку морфологической таблицы (набора морфологических классов, атрибутов, типов парадигм), что обеспечивает мультиязычность, а также возможность ее согласования с морфоанализаторами сторонних производителей.

2.2. Система Diglex

Система Diglex предназначена для описания и поиска в тексте шаблонных лексических конструкций. Язык описания шаблонов [Жигалов и др., 2002] позволяет определять произвольные символьные выражения, указывать альтернативные выражения, образуемые при использовании сокращений, аббревиатур, синонимов, пропусков и перестановок в лексическом составе конструкции, задавать условия, опциональные подвыражения, дистантный контекст и т.д. Язык Diglex DSL также позволяет описывать структуру слова за счет введения ограничений на его части (окончания, приставки и т.п.) – ограничение длины и задание списка допустимых значений.

Для реализации редактора языка Diglex DSL (рисунок 3) был выбран подход на основе концепции «проекционного редактора». Подход является симбиозом представления языка в виде текста и набора графических примитивов, и имеет ряд преимуществ по сравнению с текстовым представлением.

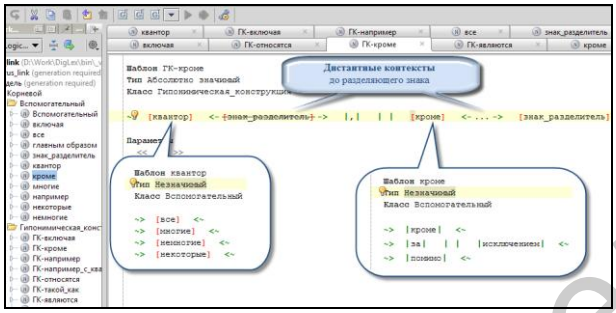

Рисунок 3– Проективный редактор словаря Diglex

Редактор обладает многими возможностями современных сред разработки: «автодополнение», интеграция с системами контроля версий, поиск и подсвечивание ошибок на этапе написания кода, расширенная навигация по коду, различные упрощающие действия и т.п.

Система снабжена модулем тестирования, который позволяет обработать фрагмент текста и визуализировать результат, сгруппированный по шаблонам и классам шаблонов.

2.3. Система SemanticVoc

Система SemanticVoc предназначена для создания семантико-синтаксических словарей, которые ограничивают синтаксическую сочетаемость и проверяют согласованность грамматических и семантических признаков терминов (вершин синтаксических групп) в соответствии с правилами согласования и управления. Словарь включает описание семантико-синтаксических моделей в виде древовидной иерархической структуры (рисунок 4), вершиной которой является лексическая метка (идентификатор модели), на следующем уровне перечисляются актанты, характеризующие соответствующую валентность, а каждый актант описывается набором семантических и грамматических характеристик, которые являются ограничениями для зависимых слов. Каждая модель может быть приписана любому количеству словоформ, лексем или обобщенных лексем, т.е. лексем, описанных в терминах грамматических и семантических категорий без указания нормальной формы.

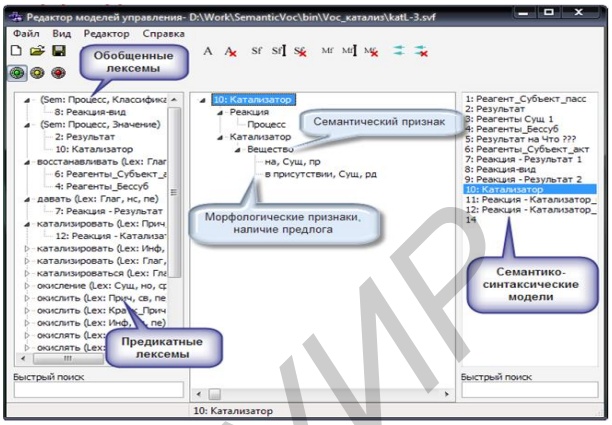

Рисунок 4 – Редактор семантико-синтаксических моделей

Предложенная структура семантико-синтаксические моделей предоставляет широкие возможности моделирования языковых связей в тексте. Так, модель может не содержать синтаксических ограничений и представлять собой онтологические отношения ПО, или модель может описываться без семантических характеристик и соответствовать классическим моделям управления. Обобщение моделей позволяет компактно определить многие языковые конструкции, варианты взаимосвязи слов в предложении и различные словарные группы.

Важным свойством системы, на наш взгляд, является редактируемая и настраиваемая система семантических признаков (в зависимости от ПО) и морфологическая модель языка, что, в частности, обеспечивает условную независимость от языка.

Созданная система является универсальным средством, реализующим словарь семантико-синтаксических шаблонов, и может использоваться в системах, обрабатывающих связный текст, для широкого круга задач. Ядро компонента представляет собой отдельную библиотеку, которая обеспечивает полный набор функций по работе со словарем, а также дополнительные сервисные функции поиска соответствующего актанта, проверки управления или согласования для двух входящих абстрактных терминов. Компонент позволяет создавать независимые xml-словари, а также согласовывать словарь с терминологическим автоматического словарем, созданным с помощью словарного компонента KLAN.

3. Инструмент создания корпусов

Инструментарий, поддерживающий создание и изучение текстовых корпусов должен включать средства предварительной разметки текста, поиска данных в корпусе, получения статистической информации и предоставления результатов

пользователю в удобной форме. Данный функционал объединен в специализированной системе для работы с корпусами – корпус-менеджере [Захаров и др., 2011].

На текущий момент создан один из важнейших компонентов корпус-менеджера - модуль разметки текста MarkSystem. Система разметки текстов позволяет приписывать фрагментам текста различные лингвистические признаки. В качестве фрагмента может выступать слово, неразрывная цепочка слов (связный фрагмент) или множество неразрывных цепочек, не образующих связный фрагмент (разрывный фрагмент).

Помимо лингвистической разметки данная система позволяет производить семантическую разметку. Семантическая разметка предметно ориентирована, поскольку определяется онтологией ПО и делится на два типа:

- терминологическая разметка, которая в первую очередь предназначена для фиксации в тексте имен понятий ПО,
- разметка отношений (или ситуаций, представляющих собой многоместные отношения), в которых размеченные сущности выступают в определенных семантических ролях.

3.1. Конкорданс

Конкорданс – традиционный способ изучения корпуса текста. Создание конкордансов рассматривается как первый этап в работе по составлению словарей. Конкорданс, предоставляя (многочисленные) контексты употребления слов, позволяет выделить основные значения слова по его сочетаемости, месту в синтаксической структуре предложения и т.д. Чем больше параметров фильтрации поддерживает инструмент построения конкорданса (например, по лемме или словосочетанию, по индексу частотности или количеству словоупотреблений, по части речи или синтаксической/семантической структуре), чем разнообразнее анализируемые параметры, тем лучше такой конкорданс может быть использован для систематизации информации и формирования лингвистических ресурсов [Добрынина, 2012].

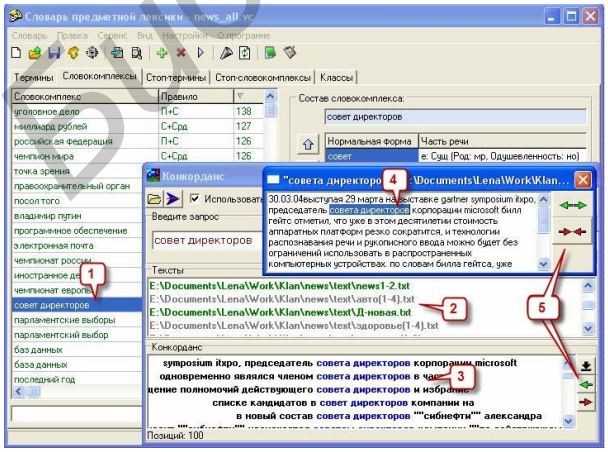

Рисунок 5 – Терминологический конкорданс

На рисунке 5 продемонстрирован пользовательский интерфейс, визуализирующий конкорданс для выбранного многословного термина словаря KLAN.

Для любого однословного или многословного термина (1) можно вызвать модуль построения конкорданса, который осуществит в реальном времени поиск в текстах (2), расположенных в заданной директории. Результатом будут все короткие контексты термина (3), обнаруженные в корпусе. При просмотре контекстов вхождений терминов пользователь может самостоятельно определять длину просматриваемых фрагментов (5) (поддерживается пословное расширение контекста) или перейти к просмотру широкого контекста конкретного вхождения термина (4), который также может быть расширен (5).

Терминологический конкорданс дает полный индекс терминов в ближайших и расширенных контекстах. Таким образом, конкорданс осуществляет обратную связь словаря, словарных терминов с корпусом и обеспечивает своего рода лингвистическую разметку на морфологическом и поверхностно-синтаксическом уровне.

На данный момент модуль конкорданса взаимодействует только с системой KLAN. В дальнейшем, планируется внедрить данный модуль в корпус-менеджер для того, чтобы можно было формировать конкорданс по любому основанию, используемому при аннотировании текстов.

Заключение

В работе представлен комплексный программный инструментарий для разработки лингвистических ресурсов. Каждый инструмент может использоваться независимо от других или в составе любой конфигурации из имеющихся модулей.

Дальнейшие исследования будут направлены на развитие методов и средств обучения и автоматического формирования лингвистических ресурсов (в первую очередь, семантико-синтаксических словарей) на основе обучающего корпуса текстов. Также будут развиваться средства поиска в корпусах, визуализации информации из различных текстов в виде совокупности контекстов фактов, которое будет служить основой для проведения исследований семантических свойств текста.

Работа выполняется при финансовой поддержке РФФИ (проект № 12-07-31216 «Разработка методов создания информационной системы, сочетающей семантическое и текстовое представление информации») и Президиума РАН (Интеграционный проект СО РАН № 15/10 «Математические и методологические аспекты интеллектуальных информационных систем»)

Библиографический список

[Богданова, 2010] Богданова, С.Ю. Исследование слова и предложения компьютерными методами: цели и перспективы / С. Ю. Богданова // Слово в предложении: кол. монография / под ред. Л.М. Ковалевой (отв. ред.), С.Ю. Богдановой, Т.И. Семеновой. –Иркутск: ИГЛУ, 2010. – С.194-214.

[Большаков, 2002] Большаков, И.А. Какие словосочетания следует хранить в словарях? / И. А. Большаков // Труды международного семинара Диалог'2002 по компьютерной лингвистике и ее приложениям. Протвино: 2002. Т.2. С.61–69.

[Большакова и др., 2010] Большакова, Е.И. Система для поиска и выделения конструкций в тексте на естественном языке / Е. И. Большакова, А. А. Носков // Труды 12-й национальной конференции по искусственному интеллекту с международным участием – КИИ-2010. – Москва: Физматлит, 2010. Т.3. -С.137-145.

[Волкова и др., 1998] Волкова, И.А. Компьютерный словарь моделей управления русских глаголов (экспериментальный вариант) / И.А. Волкова, И.Г. Головин, О.Ф. Кривнова// Труды Международного семинара по компьютерной лингвистике и ее приложениям "Диалог'98" / под ред. А.С. Нариньяни. –Казань: Хэтер, 1998.

[Гаврилова и др., 2001] Гаврилова, Т.А. Базы знаний интеллектуальных систем / Т.А. Гаврилова, В.Ф. Хорошевский // Учебник. СПб.:Питер, 2001.

[Добрынина, 2012] Добрынина, К.С. О методике работы над конкордансами / К.С. Добрынина // Филологические науки. Вопросы теории и практики. –Тамбов: Грамота, 2012. № 1 (12). С. 59-63.

[Жигалов и др., 2002] Жигалов, В.А. Система ALEX как средство для многоцелевой автоматизированной обработки текстов / В. А. Жигалов, Д. В. Жигалов, А. А. Жуков, И. С. Кононенко, Е. Г. Соколова, С. Ю. Толдова // Труды международного семинара Диалог'2002 по компьютерной лингвистике и ее приложениям. Т.2. –М.: Наука, 2002. –С.192–208.

[Загорулько и др., 2009] Загорулько, Ю.А. Технология анализа документов в информационных системах поддержки научной и производственной деятельности / Ю. А. Загорулько, Е. А. Сидорова // Автометрия, 2009. Т.45, №6. –С. 38-45.

[Захаров и др., 2011] Захаров, В.П. Корпусная лингвистика / В.П. Захаров, С.Ю Богданова // Учебник для студентов гуманитарных вузов. – Иркутск: ИГЛУ, 2011. – 161 с.

[Лукашевич, 2011] Лукашевич, Н.В. Тезаурусы в задачах информационного поиска / Н.В. Лукашевич// –М.: Издательство Московского университета, 2011. –512 с.

[Лукашевич и др., 2008] Лукашевич, Н.В. Отбор словосочетаний для словаря системы автоматической обработки текстов / Н.В. Лукашевич, Б.В. Добров, Д.С. Чуйко // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2008». Вып. 7 (14). М.: РГГУ, 2008. С. 339–344.

TOOLS FOR LANGUAGE RESOURCES SOFTWARE DEVELOPMENT

Zagorulko M.Yu., Sidorova E.A.

A.P. Ershov Institute of Informatics Systems Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

zagorulko_maxim@yahoo.com

lena@iis.nsk.su

The paper presents software tools designed for linguistic research and development of linguistic resources: text corpora and domain-specific dictionaries. Technology of dictionaries creation is focused on subject terminology extraction, as well as lexical-semantic and syntactic and semantic structures are proposed.

Introduction

The development of natural language interfaces for users of information systems providing with more convenient means of access, search and text presentation call for solution of different linguistic tasks. The solution of these tasks requires certain linguistic resources, the creation of which (in their absence) is a separate task. Software tools designed for linguistic research and development of linguistic resources such as text corpora and domain-specific dictionaries is considered.

Main Part

During development of natural-language modules different linguistic tasks grow up and their solution requires the use of a variety of linguistic resources. In our process chain, we consider the following types of linguistic resources: text corpora, generic and subject dictionaries, semantic dictionary, document model, fact scheme.

Our tools support the creation of several types of dictionaries: terminological dictionary, lexical and semantic dictionaries, semantic and syntactic dictionaries. All systems are equipped with a custom dictionary editor of the resource, testing module and software API, which allows using the developed linguist expert resource offline.

Tools that support the creation and study of text corpora should include a means of preliminary marking of text, data retrieval, statistical information obtaining and providing the results to the user in a convenient form. This functionality is integrated in a specialized system to work with text corpora - the corpus-manager

Conclusion

This paper presents software tools for development of linguistic resources. Each tool can be used independently or as part of any configuration of the available modules. Further research will be aimed at methods and tools developing for learning and automatic creation of linguistic resources through the training corpus. Also we plan to develop different tools to study the semantic properties of the text.