УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ
«МИНСКИЙ ГОСУДАРСТВЕННЫЙ
ЛИНГВИСТИЧЕСКИЙ УНИВЕРСИТЕТ»

УДК 811.521’33 (043.3)

КРАВЧЕНКО
Сергей Юрьевич

СИСТЕМА БАЗОВОГО АВТОМАТИЧЕСКОГО ЛИНГВИСТИЧЕСКОГО АНАЛИЗА ТЕКСТОВ ЯПОНСКОГО ЯЗЫКА

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата филологических наук

по специальности 10.02.21 – прикладная и математическая лингвистика

Минск, 2014

Научная работа выполнена в учреждении образования «Минский государственный лингвистический университет» на кафедре информатики и прикладной лингвистики

Научный руководитель: **Зубов Александр Васильевич,**
доктор филологических наук, профессор, заведующий кафедрой информатики и прикладной лингвистики Минского государственного лингвистического университета

Официальные оппоненты: **Совпель Игорь Васильевич,**
доктор технических наук, профессор, профессор кафедры информационных систем управления Белорусского государственного университета

Кострова Мария Алексеевна,
кандидат филологических наук, ст. преподаватель кафедры восточных и европейских языков Нижегородского государственного лингвистического университета, Россия

Оппонирующая организация: **УО «Гродненский государственный университет имени Янки Купалы»**

Защита состоится 19 декабря 2014 г. в 14:00 на заседании совета по защите диссертаций К 02.22.02 в учреждении образования «Минский государственный лингвистический университет» по адресу: 220034, г. Минск, ул. Захарова, 21, E-mail: info@mslu.by, тел. ученого секретаря: (017) 284-81-56.

С диссертацией можно ознакомиться в библиотеке учреждения образования «Минский государственный лингвистический университет».

Автореферат разослан ____ ноября 2014 г.

Ученый секретарь
совета по защите диссертаций Р.В. Детскина

ВВЕДЕНИЕ

Поскольку естественный язык (ЕЯ) является универсальным средством описания действительности и коммуникации с вычислительной системой, то актуальность его автоматической обработки в составе современных информационных технологий, безусловно, очень высока. Данное направление, называемое иначе NLP (Natural Language Processing), связано, чаще всего, с моделированием и обработкой ЕЯ в целях автоматизации его понимания. Учитывая, что ЕЯ «проявляется» как в виде текста, так и в виде речи, будем говорить в дальнейшем только о тексте (в самом широком смысле этого слова). Суть понимания текста состоит в представлении его содержания в терминах и отношениях некоторой заданной системы знаний, определяемой целевой задачей, например, это может быть множество ключевых слов с указанием для них синонимических и иерархических отношений, множество объектов с указанием для них атрибутивных и функциональных отношений. В любом случае автоматизация понимания текста требует разработки процедур его лингвистического анализа. Об этих процедурах, а именно, лексического, лексико-грамматического, синтаксического и семантико-синтаксического анализа, характерных для большинства важнейших приложений NLP (автоматизации инженерии знаний, информационного поиска, машинного перевода, автоматического реферирования и других), в совокупности говорят как о базовом лингвистическом процессоре (БЛП).

К настоящему времени такие БЛП уже разработаны для немецкого, английского, французского и других ЕЯ. При этом особое внимание уделяется достижению их высоких качественных показателей, что особенно важно с точки зрения приложений. Что касается японского языка, то он относится к классу иероглифических ЕЯ и уже в силу этого существенно отличается от упомянутых выше. Существующие методы автоматического анализа текстов на этом языке при всём своём алгоритмическом многообразии являются в основном проблемно-ориентированными, использующими сложные иерархические и многоуровневые системы классификации для описания содержания текстов и, как следствие этого, являются несводимыми воедино в целях БЛП. Поэтому с точки зрения его функциональности необходим системный подход к решению проблемы и разработка, в этом плане, с учётом особенностей японского языка как объекта моделирования, новых методов и лингвистических ресурсов, в том числе, классификаторов его свойств на различных уровнях глубины языка, требуемых множеств лингвистических правил и соответствующих алгоритмов. Промышленный характер БЛП требует, чтобы он обладал высокой точностью и скоростью. Такие показатели не могут быть получены только активно развиваемыми вероятностно-статистическими методами, которые довольно ограничены с точки зрения их дальнейшего улучшения и, к тому же, требуют больших по объему и, как правило,

создаваемых вручную, лингвистических ресурсов для обучения алгоритмов. Основной упор должен быть сделан на разработку и использование лингвистических правил анализа текста, уже показавших свою эффективность при построении БЛП для других ЕЯ. Таким образом, развитая лингвистическая база знаний (ЛБЗ) БЛП в совокупности с эффективными технологиями её построения и тестирования всех функциональных модулей БЛП является основным ресурсом достижения его высоких качественных показателей.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Связь работы с крупными научными программами и темами

Диссертационное исследование выполнялось на кафедре информатики и прикладной лингвистики МГЛУ в рамках госбюджетной темы НИР «Создание параллельных (англо-русского и франко-русского) корпусов тегированных разножанровых текстов и их использование для совершенствования учебного процесса и научной деятельности» в рамках комплексной программы «Лингвистика и образование» (государственная программа фундаментальных исследований «Непрерывное образование») на 2001–2004 гг., а также в рамках программы научных исследований отдела разработки средств интеллектуализации информационных систем иностранного унитарного предприятия «АйЭйчЭс Глобал».

Цель и задачи исследования

Целью диссертационной работы является разработка лингвистического, алгоритмического и технологического обеспечения базового лингвистического процессора японского языка и его реализация в виде промышленного прототипа. Для достижения поставленной цели необходимо решить следующие основные задачи:

1) выявить особенности японского языка как объекта моделирования, наиболее существенно влияющие на эффективность решения задач его автоматического анализа;
2) сформулировать концепцию БЛП для японского языка, ориентированную на промышленный характер решаемых задач, и разработать его структурно-функциональную схему;
3) разработать формат базового словаря БЛП и классификаторы свойств японского языка на лексико-грамматическом, синтаксическом и семантико-синтаксическом уровнях, определить эффективные средства формального описания словарных статей базового словаря (БС) и лингвистических правил анализа текста;
4) построить технологии аннотирования и осуществить разработку базового словаря и базового корпуса текстов (БКТ) БЛП, а также построить технологию тестирования всех функциональных модулей БЛП с целью получения его качественных характеристик;

5) разработать множества лингвистических правил и основанные на них эффективные алгоритмы лексического, лексико-грамматического, синтаксического и семантико-синтаксического анализа текстов японского языка, обеспечивающие в совокупности функциональность БЛП;
6) построить прототип БЛП японского языка, определить на основе тестирования его качественные характеристики и внедрить в промышленную эксплуатацию.

Объектом исследования являются процессы автоматического лингвистического анализа текстов японского языка на различных уровнях его глубины.

Предметом исследования являются методы, алгоритмы и лингвистические ресурсы базового автоматического лингвистического анализа текстов японского языка.

Выбор объекта и предмета исследования обусловлен необходимостью совершенствования существующих систем автоматического лингвистического анализа текстовых документов.

Материалом исследования являются четыре корпуса текстов:

- корпус текстов на японском языке, состоящий преимущественно из текстов научного стиля объёмом 33 тыс. предложений, аннотированных лексико-грамматической информацией;
- корпус текстов на японском языке, представляющий собой выборку текстов из научных и научно-популярных статей, объёмом 1000 предложений с однозначно выделенными границами предложений;
- корпус текстов на японском языке, представляющий собой выборку текстов из газетных статей, научно-популярных изданий и научных патентов, объёмом 3 тыс. предложений, аннотированных лексико-грамматической информацией;
- корпус текстов на японском языке, представляющий собой выборку текстов из научных статей и патентов, объёмом 100 предложений, аннотированных лексико-грамматической и синтаксической информацией.

Положения, выносимые на защиту

1. Концепция базового ЛП японского языка, которая в отличие от существующих ЛП является универсальной по отношению к решаемым задачам автоматической обработки текста, ориентированной на их промышленный характер, и учитывает особенности японского языка как объекта моделирования, а также структурно-функциональная схема БЛП, согласно которой основные компоненты его ЛБЗ организованы в виде отдельных модулей, соответствующих функциональным уровням анализа текста.

2. Формат базового словаря японского языка, ориентированный на использование минимальных лексических единиц (МЛЕ) в качестве его словарных статей, а также их фонетической, графической и смысловой эквивалентности и словоизменительных свойств лексических единиц, которые в совокупности могут быть описаны на языке расширенных регулярных выражений, что существенно повышает производительность процесса построения БС и его обработки в составе БЛП.
3. Технология аннотирования базового корпуса текстов, основанная на сформулированных принципах определения границ МЛЕ и их лексико-грамматических кодов (ЛГК), на итеративном характере процесса, позволяющем постепенно переходить от его ручных процедур к автоматическим, а также оценка объема БКТ, являющегося достаточным для построения качественных вероятностных алгоритмов лексико-грамматического анализа текстов японского языка.
4. Новое в совокупности лингвистическое обеспечение БЛП японского языка, включая классификаторы его свойств на лексико-грамматическом, синтаксическом и семантико-синтаксическом уровнях, базовый корпус текстов, базовый словарь и множества лингвистических правил нормализации текста, распознавания границ предложений, определения ЛГК неизвестных по отношению к базовому словарю МЛЕ, корректировки ЛГК МЛЕ, синтаксического и семантико-синтаксического анализа текста, а также формализация лингвистических правил на языке расширенных регулярных выражений и принципиальные схемы их обработки на соответствующих этапах анализа текста.
5. Алгоритмы нормализации текста и его графемного анализа, алгоритмы морфологического анализа текста и последующей корректировки его результатов, алгоритмы синтаксического и семантико-синтаксического анализа текстов, а также технология тестирования всех функциональных модулей БЛП, основанная на использовании эталонных корпусов текстов.
6. Прототип промышленного БЛП японского языка, его качественные характеристики и результаты внедрения для решения актуальных прикладных задач.

Личный вклад соискателя

Диссертационное исследование является самостоятельной работой автора. Все вошедшие в диссертацию результаты получены при непосредственном личном участии автора. Научному руководителю принадлежит выбор направления исследований и обсуждения результатов.

Апробация результатов диссертации

Результаты исследований, включенные в диссертацию, докладывались и обсуждались на: заседании кафедры информатики и прикладной лингвистики МГЛУ в 2014 г.; Международном конгрессе по информатике: информационные системы и технологии CSIST'2011 (Минск, Беларусь, 2011); Международном конгрессе по информатике: информационные системы и технологии CSIST'2013 (Минск, Беларусь, 2013); ежегодной конференции преподавателей и аспирантов МГЛУ (Минск, Беларусь, 2012); 24-й Международной конференции по компьютерной лингвистике COLING'2012 (Мумбаи, Индия, 2012).

Разработанный прототип базового лингвистического процессора японского языка внедрён в состав промышленного многоязычного лингвистического процессора известной системы инженерии и управления знаниями IHS Goldfire, что подтверждается актом о практическом использовании результатов исследования.

Опубликованность результатов диссертации

По теме диссертации опубликовано 9 научных работ, среди которых 5 опубликованы в научных журналах, включенных в Перечень научных изданий Республики Беларусь, утвержденный Высшей аттестационной комиссией (объемом 2,3 авторского листа), 3 – в сборниках материалов научных конференций (объемом 1,1 авторского листа), 1 – в сборнике научных статей (объемом 0,4 авторского листа). Общий объем опубликованных материалов составляет 3,8 авторского листа.

Структура и объём диссертации

Диссертация включает перечень условных обозначений, введение, общую характеристику работы, основную часть работы, заключение и библиографический список, включающий 116 наименований на русском, английском, немецком и японском языках (из них 9 публикаций соискателя).

Основная часть состоит из трех глав. Необходимость решения поставленных в диссертации задач определила следующую последовательность изложения. В первой главе рассматривается общая функциональность базового лингвистического процессора, исследуются особенности японского языка как объекта моделирования, а также анализируются существующие подходы к решению задач БЛП японского языка. Во второй главе предлагается принципиальная схема БЛП японского языка, а также описываются технологии создания основных компонентов его лингвистической базы знаний. В третьей главе представлены принципиальные алгоритмы работы основных функциональных модулей БЛП, осуществляется анализ результатов работы предложенного прототипа, а также способы их применения.

Полный объем диссертации составляет 152 страниц, в том числе основной текст – 132 страницы, 9 таблиц и 17 рисунков – 17 страниц, и библиографический список – 10 страниц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы диссертации, определены цели и задачи исследования.

В первой главе показаны актуальность задач автоматической обработки текстов естественного языка и лингвистических процессоров как основы их решения, а также функциональность базового лингвистического процессора и особенности японского языка как объекта моделирования в целях БЛП. Дан анализ существующих подходов к решению задач БЛП японского языка.

Постоянно возрастающая роль текста как основного источника знаний о предметной области/внешнем мире требует решения актуальной задачи автоматизации его обработки в составе современных информационных технологий. И оно может быть получено на основе развитых лингвистических процессоров для различных ЕЯ, в том числе и для японского языка. Учитывая, что большинству основных задач автоматической обработки текста характерно наличие общих процедур, то об их совокупности можно говорить как о базовом лингвистическом процессоре. Его функциональность включает:

- графемный анализ, на котором формализуются границы основных единиц анализа в тексте, таких как абзацы, предложения, словосочетания и слова;
- лексико-грамматический анализ, на котором моделируются лексико-грамматические характеристики единиц анализа;
- синтаксический анализ, на котором формализуются синтаксические связи между словами в предложении;
- семантико-синтаксический анализ, который используется для моделирования семантических характеристик синтаксических связей между единицами текста.

Таким образом, БЛП представляет собой транслятор, преобразующий текстовую информацию, оформленную средствами ЕЯ, в некоторое формализованное представление на одном или нескольких из выбранных уровней представления содержания текста. Переход к формализованному представлению, как правило, осуществляется постадийно, что находит отражение в модульной организации БЛП.

Функциональность БЛП обеспечивается его лингвистической базой знаний, которая включает четыре основных компонента: классификатор (систему формализмов) для описания содержания текста на лексико-грамматическом, синтаксическом и семантико-синтаксическом уровнях глубины языка; базовый корпус текстов, размеченный в соответствии с классификатором свойств ЕЯ и предназначенный для обучения вероятностно-статистических моделей анализа

текста, тестирования лингвистических гипотез и отдельных модулей БЛП; базовый, размеченный лексико-грамматическими классами слов, словарь ЕЯ; множество лингвистических правил (шаблонов, паттернов) анализа текста на различных уровнях глубины ЕЯ, разработанных экспертами-лингвистами. Классификатор свойств ЕЯ является ключевым компонентом ЛБЗ, определяющим форматы базового словаря, базового корпуса текстов и лингвистических правил, а также качественные характеристики БЛП, которые в дополнение могут быть существенно улучшены за счёт использования преимуществ модульного подхода к проектированию как функциональности БЛП, так и его ЛБЗ.

Разработка БЛП требует учёта особенностей каждого конкретного ЕЯ, как объекта моделирования. Показано, что таковыми для японского языка являются: использование одновременно нескольких видов письменности в текстах – катаканы, хираганы, кандзи и ромадзи; отсутствие пробелов или других формальных показателей для обозначения границ слов, записанных с использованием катаканы, хираганы и кандзи; использование развитой системы суффиксов, служебных слов и вспомогательных глаголов для передачи лексико-грамматических значений слов; специфический порядок слов в предложении; эллипсис (случаи опущения личных местоимений, а также отсутствия глаголов в предикативных группах); тематическое выделение членов предложения; неоднозначность интерпретации синтаксических связей в сложных словах, предикативных группах и сложных предложениях; аналитический характер семантической организации высказываний.

Анализ существующих методов решения задач автоматического анализа текстов японского языка показал, что при всём своём алгоритмическом многообразии они являются в основном проблемно-ориентированными, использующими сложные иерархические и многоуровневые системы классификации для описания содержания текстов и, как следствие этого – являются несводимыми воедино в целях БЛП. Все сказанное ставит задачу системного подхода к разработке новых методов и лингвистических ресурсов, в частности – классификаторов свойств японского языка и требуемого множества лингвистических правил. Создаваемый БЛП должен обладать высокой точностью и скоростью. В отличие от активно развиваемых вероятностно-статистических методов основной упор в данном исследовании сделан на разработку и использование лингвистических правил анализа текста, уже показавших свою эффективность при построении БЛП целого ряда ЕЯ.

Во второй главе рассмотрен круг вопросов, касающихся комплексного решения задачи построения лингвистической базы знаний БЛП японского языка. Разработана и обоснована, исходя из особенностей японского языка, структурно-функциональная схема его БЛП, согласно которой основные компоненты ЛБЗ организованы в виде отдельных модулей, соответствующих функциональным уровням обработки текста (рисунок).

Разработанный лексико-грамматический классификатор японского языка включает 143 лексико-грамматических кода, в основу которого обоснованно положен морфемный подход. Существенным отличием предложенного классификатора является его практическая ориентация, что нашло отражение в выборе способа сегментации целого ряда минимальных лексических единиц (минимальных значимых единиц анализа текста, таких как морфемы, знаки препинания, элементы формул и др.) и их ЛГК. Было предложено рассматривать два типа морфем – корневые и аффиксальные. Все классы корневых морфем в построенном классификаторе можно условно разбить на несколько групп – корни существительных (4 ЛГК), корни прилагательных (3 ЛГК), корни глаголов (13 ЛГК), корни наречий (1 ЛГК), корни местоимений (8 ЛГК), корни числительных (1 ЛГК), междометия (1 ЛГК), союзы (3 ЛГК) и специальные МЛЕ (15 ЛГК). Аффиксальные морфемы можно разделить на следующие подклассы — суффиксы форм и спряжений глаголов и прилагательных (58 ЛГК), словообразующие суффиксы (12 ЛГК), семантические аффиксы (4 ЛГК), аффиксы числительных (3 ЛГК), показатели падежей (11 ЛГК), именные послелоги (1 ЛГК), аффиксы вежливой речи (2 ЛГК) и частицы (3 ЛГК).

Предложенная классификация морфем позволяет эффективно решать задачи обработки в текстах слов-заимствований по принципу «катакана-ромадзи» и аббревиатур. Она существенно ограничивает количество вариантов морфологического и синтаксического разбора предложений за счёт более точного описания всевозможных комбинаций корневых и аффиксальных морфем без усложнения структуры компонентов ЛБЗ, значительно расширяет описательные возможности лингвистических правил анализа текста и снижает синтаксическую многозначность анализируемых предложений. Определённую гибкость данному классификатору обеспечивает разработанная дополнительно система макро-кодов, включающая 414 кодов. Она позволяет группировать ЛГК в более компактные и наглядные формы записи, что очень важно при разработке лингвистических правил анализа текста.

В основу построенного синтаксического классификатора положено понятие минимальной синтаксической единицы (МСЕ) как корневой морфемы со всеми относящимися к ней аффиксами и другими корневыми морфемами в случае сложных слов, что обеспечило классификацию МСЕ на основании лексико-грамматических характеристик входящих в их состав МЛЕ. Полный список МСЕ включил 90 классов.

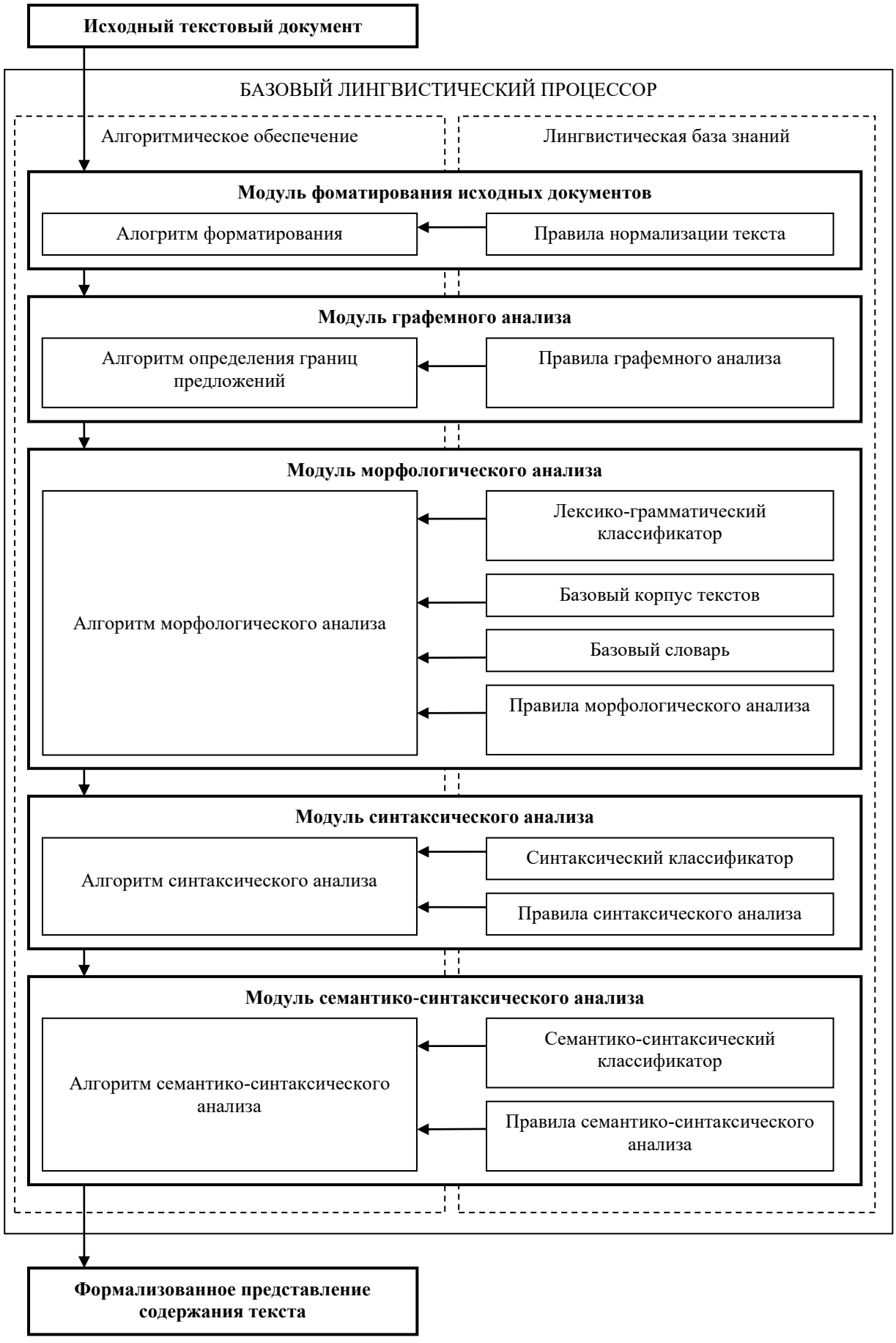

Рисунок – Общая схема БЛП японского языка

Для классификации синтаксических связей между МСЕ с целью описания подчинительных и сочинительных отношений в предложении определены специальные классы синтаксических групп – классы расширенных синтаксических групп (РСГ), обладающие свойством рекурсивности (например, классы именных групп, расширенных групп сказуемого, подчинённых предложений с глагольной группой сказуемого и др.) Полный список РСГ включил 46 классов. При этом описательные возможности синтаксического классификатора также расширены за счёт построенной системы макро-кодов, которая включила 132 кода.

Учитывая целевую установку БЛП на этапе семантико-синтаксического анализа, в качестве выходного результата на данном этапе предложено использовать предикативные группы. Для их описания в виде 28-компонентной структуры используется классификация именных аргументных групп на основе поверхностных падежей их главных слов и классификация сказуемых на основе классов слов, которыми они выражены:

r__Predicate

- NoCase_Object - именное сказуемое или дополнение без падежного показателя,
- CHA_Object - доп. в именительном тематическом падеже,
- CHA_Suffix - показатель им. тематического падежа,
- CGA_Object - доп. в именительном рематическом падеже,
- CGA_Suffix - показатель им. рематического падежа,
- CWO_Object - доп. в винительном падеже,
- CWO_Suffix - показатель винительного падежа,
- CNO_Object - доп. в родительном падеже,
- CNO_Suffix - показатель родительного падежа,
- CNI_Object - доп. в дательном падеже,
- CNI_Suffix - показатель дательного падежа,
- CDE_Object - доп. в творительном падеже,
- CDE_Suffix - показатель творительного падежа,
- CTO_Object - доп. в совместном падеже,
- CTO_Suffix - показатель совместного падежа,
- CHE_Object - доп. в направительном падеже,
- CHE_Suffix - показатель направительного падежа,
- CKARA_Object - доп. в исходном падеже,
- CKARA_Suffix - показатель исходного падежа,
- CYORI_Object - доп. в исходно-сравнительном падеже,
- CYORI_Suffix - показатель исходно-сравнительного падежа,
- CMADE_Object - доп. в предельном падеже,
- CMADE_Suffix - показатель предельного падежа,
- Adjective - адъективное сказуемое или обстоятельство,
- Adverb - адвербиальное обстоятельство,

- Verb - глагольное сказуемое,
- Conjunction - подчинительный союз (в случае придаточного предложения),
- HeadWord - главная СГ из подчиняющего предикатива.

В интересах приложений БЛП, особенно с точки зрения задач автоматизации инженерии знаний, вопросно-ответной функциональности, семантико-синтаксический классификатор дополнен конструкцией, соответствующей простой именной группе.

r__SimpleNounPhrase
- NounPhrase

Построена технология аннотирования базового корпуса текстов, основанная на сформулированных принципах определения границ МЛЕ и их ЛГК, а также на итеративном характере процесса, позволяющем постепенно переходить от его ручных процедур к автоматическим.

Разработанный БКТ включил более 33000 аннотированных ЛГК предложений японского языка из текстов патентов, газетных и журнальных статей, пользовательских запросов и художественной литературы. Показано, что данный объём БКТ является достаточным для построения качественных вероятностных алгоритмов морфологического анализа текстов японского языка.

Разработан формат базового словаря японского языка, ориентированный на использование МЛЕ в качестве его словарных статей, а также их фонетической, графической и смысловой эквивалентности и словоизменительных свойств лексических единиц, которые в совокупности могут быть описаны на языке расширенных регулярных выражений, что существенно повышает производительность процесса построения БС и его обработки в составе БЛП. Например, лексико-грамматическая характеристика МЛЕ 太陽 (ТАЙЁ:, *солнце*) представлена в БС следующим образом:

太陽 NN NP

где помимо самой описываемой МЛЕ (太陽) через табуляцию указаны все возможные ЛГК для данной МЛЕ (NN – ЛГК корня существительного, NP – ЛГК корня имени собственного).

Информация о фонетической, графической и семантической эквивалентности приводится в словарных статьях специальных разделов БС, выделенных фигурными скобками и словами ФОНЕТИЧЕСКАЯ, ГРАФИЧЕСКАЯ и СЕМАНТИЧЕСКАЯ соответственно:

ФОНЕТИЧЕСКАЯ
{
    太陽/NN → たいよう

}

ГРАФИЧЕСКАЯ
{
	太陽/NN = たいよう/NN = タイヨウ/NN
}

СЕМАНТИЧЕСКАЯ
{
	太陽/NN = お日様/NN = 日輪/NN
}

где в фигурных скобках записаны словарные статьи с информацией об эквивалентности соответствующего типа. Так, например, словарная статья в разделе ФОНЕТИЧЕСКАЯ читается как «МЛЕ 太陽 с ЛГК NN произносится как た い よ う (ТАЙЁ:)»; словарная статья в разделе ГРАФИЧЕСКАЯ обозначает «МЛЕ 太陽 с ЛГК NN имеет вариант написания в виде МЛЕ たいよう, а также вариант написания в виде МЛЕ タイヨウ, причём каждый из данных вариантов с соответствующими ЛГК также имеет варианты написания в виде двух других»; словарная статья в разделе СЕМАНТИЧЕСКАЯ читается как «МЛЕ 太陽 с ЛГК NN, МЛЕ お日様 с ЛГК NN и МЛЕ 日輪 с ЛГК NN обозначают один и тот же предмет».

Информация о словоизменительных характеристиках слов фиксируется в словарных статьях соответствующих разделов БС:

ОТРИЦАНИЕ_ГЛАГОЛА
{
	<1 VB1S > SV1SU → $1 さ/SV1SA な/SPNEG い/SPI
}

Словарная статья в данном разделе обозначает следующее: «если начальная (словарная) форма глагола совпадает с шаблоном VB1S SV1SU (последовательности МЛЕ с ЛГК VB1S SV1SU), то отрицательная форма должна быть сформирована за счёт присоединения к корневой морфеме последовательности МЛЕ – さ/SV1SA な/SPNEG い/SPI».

Семантические признаки МЛЕ фиксируются в виде списков МЛЕ и их ЛГК в специальных разделах БС, например:

ПЕРЕХОДНЫЕ_ГЛАГОЛЫ
{
	動か/VB1S
	高め/VB2
	染め/VB2
	...
}

где МЛЕ 動か с ЛГК VB1S, МЛЕ 高め с ЛГК VB2 и МЛЕ 染め с ЛГК VB2 – корневые морфемы переходных глаголов.

Построена эффективная технология наполнения БС, в соответствии с которой он был разработан и включил 253769 уникальных МЛЕ с их ЛГК, 8542 статьи смысловых характеристик МЛЕ, 12625 статей признаков смысловой эквивалентности, 8526 статей признаков графической и фонетической эквивалентности и 3399 статей парадигм словоизменения.

Обосновано использование лингвистических правил в качестве основного механизма функциональности БЛП. На основании экспертного анализа существующих знаний о японском языке, а также построенных в результате проведённых исследований классификаторов его свойств, базового корпуса текстов, базового словаря, и с учётом функциональности БЛП, разработаны базы лингвистических правил нормализации текста (197 правил), определения границ предложений (5 правил), определения ЛГК неизвестных по отношению к БС МЛЕ (172 правила), правил корректировки ЛГК МЛЕ (297 правил), правил синтаксического анализа (69 стадий, всего 1214 правил) и правил семантико-синтаксического анализа (12 правил). Указанные лингвистические правила описаны на языке расширенных регулярных выражений. Например – правило исправления ошибочной маркировки МЛЕ:

< "おも"/JJI "い"/SPI > "だ" "す" → VB1W SV1WI

В приведённом примере слева от символа «→» - записано условие срабатывания данного правила – лингвистический шаблон в виде последовательности МЛЕ и/или соответствующих им ЛГК. Справа от символа «→» записана операция, которая должна быть осуществлена в случае совпадения данного шаблона с входными данными (в данном примере – предложения, разбитого на МЛЕ с назначенными им ЛГК). Смысл приведённого примера правила состоит в следующем: «если в предложении обнаружена последовательность МЛЕ «おも», «い», «だ», «す», в которой первым двум МЛЕ назначены ЛГК JJI и SPI, то нужно заменить ЛГК первых двух МЛЕ на ЛГК VB1W и SV1WI соответственно». При обнаружении ошибочного морфологического анализа, например, глагола おもいだす (ОМОИДАСУ, *вспоминать*), осуществлённого вероятностной моделью анализа, данное правило исправит соответствующую ошибку:

おも_JJI い_SPI だ_SP1S す_SV1SU → **おも_VB1W い_SV1WI** だ_SP1S す_SV1SU

В третьей главе рассмотрены вопросы, связанные с реализацией БЛП японского языка с использованием разработанной ЛБЗ. Разработано и представлено в виде совокупности блок-схем алгоритмическое обеспечение БЛП, построена технология тестирования всех его функциональных модулей, получены значения качественных характеристик БЛП, приведены данные о его практическом использовании.

Алгоритмическая часть БЛП представляет собой совокупность алгоритмов обработки текста на различных языковых уровнях в соответствии с общей схемой, представленной на рисунке.

Алгоритм форматирования исходных документов подразумевает выполнение следующих функций:

- преобразование исходных документов в текстовый вид;
- выделение абзацев из текстов исходных документов;
- нормализация текстовой информации (приведение различных вариантов одних и тех же символов к единому виду).

Преобразование исходных документов в текстовый вид и выделение из них абзацев является, в большей степени, проблемой форматирования исходных документов, чем лингвистического анализа, и решается при помощи специально предназначенной для этого известной программы. Что касается нормализации текстовой информации, то данная задача решается правилами нормализации текста, представленными в ЛБЗ.

После нормализации текста выделенные абзацы передаются модулю графемного анализа, результатом работы которого является набор последовательностей символов, соответствующих предложениям, выделенным из исходного текста.

Полученные предложения (последовательности символов) поступают далее на вход модуля морфологического анализа, работа которого основывается, прежде всего, на известной вероятностной модели Маркова и построенном алгоритме распознавания ЛГК неизвестных по отношению к БС МЛЕ.

Для исправления ошибок, допущенных Марковской моделью, далее применяется алгоритм, построенный на использовании правил исправления ошибочной маркировки МЛЕ. На вход данной группе правил поступает результат работы вероятностной модели – предложение в виде последовательности МЛЕ и соответствующих им ЛГК. На выходе – предложение с исправленными ошибками маркировки МЛЕ, что и является конечным результатом работы модуля морфологического анализа.

Далее, предложения обрабатываемого текста с выделенными МЛЕ и назначенными им ЛГК поступают на вход модуля синтаксического анализа, результатом работы которого является набор предложений со сформированными в нём синтаксическими группами. Результат синтаксического анализа подаётся на вход модуля семантико-синтаксического анализа. Обработка предложений в данном модуле осуществляется с использованием подхода, основанного на правилах, записанных в виде под-деревьев. Результатом являются предложения с выделенными из них семантико-синтаксическими структурами. Данный результат является одновременно конечным результатом работы предлагаемого БЛП.

В качестве примера выхода БЛП в таблице приведён полный разбор предложения «インドネシアコモドドラゴンについてまず聞かれるのは、その大きさだ。» (ИНДОНЭСИА КОМОДО ДОРАГОН НИ ЦУЙТЭ МАЗУ КИКАРЭРУ НО ВА, СОНО О:КИСА ДА, *первое, что спрашивают об индонезийском комодском варане – это его величина*), пронумерованный по строкам.

Таблица – Пример выхода (конечного результата работы) БЛП

Номер строки	**Выход БЛП**
1	w_VP_Finite(w_NP_Additional(w_VP_Finite(w_VP_Subordinated_OneArgument(w_NP_Min(w_Common_NP(w_Common_NP(インドネシア_NP コモド_FW) ドラゴン_FW)) w_CNI(に_CNI) w_Predicate_OneArgument_Continuous(つ_VB1K い て _SPTE)) w_VP_Finite(w_Adverb(ま ず _RB) w_Predicate_Common_Finite(聞 _VB1K か _SV1KA れ _SPRE る _SV2U))) w_NP_Formal(w_SUBST_NP(の _SUBST))) w_CHA(は _CHA) 、 _PMC w_VP_Finite(w_NP_Min(w_Common_NP(その_PDT w_Common_NP(w_NN(大き_JJI さ_SNNJJ)))) w_Desu_DA(だ_VBDE))) 。 _PMFS
2	r__SimpleNounPhrase
3	NounPhrase インドネシア_NP コモド_FW
4	r__SimpleNounPhrase
5	NounPhrase の_SUBST
6	r__SimpleNounPhrase
7	NounPhrase その_PDT 大き_JJI さ_SNNJJ
8	r__Predicate
9	CNI_Object インドネシア_NP コモド_FW
10	CNI_Suffix に_CNI
11	Verb つ_VB1K いて_SPTE
12	HeadWord 聞_VB1K か_SV1KA れ_SPRE る_SV2U
13	r__Predicate
14	Adverb まず_RB
15	Verb 聞_VB1K か_SV1KA れ_SPRE る_SV2U
16	HeadWord の_SUBST
17	r__Predicate
18	NoCase_Object その_PDT 大き_JJI さ_SNNJJ
19	CHA_Object の_SUBST
20	CHA_Suffix は_CHA
21	Verb だ_VBDE

Как видно из примера, результатом работы БЛП являются, во-первых, предложения исходного текста с выделенными МЛЕ и назначенными им ЛГК

(строка 1). МЛЕ разделены друг от друга как минимум одним символом пробела, а ЛГК приписаны к соответствующим МЛЕ через символ «_», например «インドネシア_NP». Кроме того, в предложении содержится информация о синтаксических группах, выделенных на этапе синтаксического анализа. Каждая синтаксическая группа обозначена круглыми скобками, перед которыми записывается условное обозначение данной группы, например «w_VP_Finite(...)». Непосредственно за строкой с предложением следует информация о семантико-синтаксических единицах, выделенных из данного предложения. Сначала выводятся все простые именные группы, обозначенные как «r__SimpleNounPhrase» (строки 2, 4, 6). Каждая из данных конструкций содержит соответствующее поле (строки 3, 5 и 7), в котором название поля «NounPhrase» отделено от его содержания символом табуляции. Непосредственно за списком именных групп следует список предикативных групп, обозначенных как «r__Predicate» (строки 8, 13, 17). Каждая из предикативных групп включает поля, содержание которых соответствует аргументным группам, предикативам и другим членам предложения, относящимся к конкретной группе (строки 9-12, 14-16 и 18-21). Так же, как и в случае с именными группами, содержание поля отделено от его названия символом табуляции. Содержание полей семантико-синтаксических единиц представлено в виде последовательностей МЛЕ с назначенными им ЛГК и при необходимости может быть дополнено информацией о порядковых номерах МЛЕ в предложении.

Оценка эффективности функциональности БЛП осуществлялась по таким известным в теории и практике систем автоматической обработки текста показателям как точность, полнота и F-мера. В этих целях построена технология тестирования, которая предполагает разработку на основе выборки определённого объёма текстов из БКТ эталонного корпуса текстов, в котором с использованием автоматических процедур и экспертной обработки точно выполнены операции анализа в соответствии с функциональностью каждого модуля БЛП. Такой эталонный корпус текстов подаётся на вход тестируемого модуля, осуществляется его соответствующий автоматический анализ (не принимая во внимание зафиксированные в нём результаты предшествующего точного анализа), полученные результаты автоматически сравниваются с зафиксированными ранее и вычисляются значения используемых показателей эффективности.

Полученные при тестировании разработанного БЛП оценки эффективности оказались сопоставимыми, а в ряде случаев – лучшими, по сравнению с известными системами, решающими отдельные задачи автоматического анализа текстов японского языка. Так, например, показатель точности модуля морфологического анализа (97,32%) существенно превзошёл показатель системы JUMAN (93,73%) и оказался сопоставим с показателем системы ChaSen (96,95%). В целом совокупные показатели качества

построенного БЛП, в том числе и по скорости обработки текста, являются лучшими среди ЛП японского языка промышленного типа. Существенным преимуществом разработанного БЛП является его модульная структура и возможность одновременного представления результатов лингвистического анализа текстов на нескольких уровнях.

Разработанный БЛП японского языка внедрён в составе промышленного многоязычного лингвистического процессора известной системы инженерии и управления знаниями IHS Goldfire, обеспечивающего эффективное решение целого ряда актуальных прикладных задач:

- автоматического распознавания в текстовых документах знаний основных типов;
- семантического тегирования текстовых документов и запросов пользователя в целях вопросно-ответной функциональности в многоязычной информационной среде с ЕЯ-интерфейсом пользователя, и в целях мониторинга пользовательского мнения относительно отдельных изделий, товаров и услуг;
- получения интегрированной оценки информативности лексических единиц текстовых документов в целях построения принципиально нового типа их рефератов, ориентированных на структуризацию информационной потребности пользователя и выбранную им тему.

Как показал анализ результатов работы БЛП, предложенный прототип способен полностью реализовать базовую функциональность лингвистического процессора японского языка. Качество полученных результатов позволяет говорить о возможности использования данного БЛП для решения задач автоматического анализа текста в рамках японского языка. Исследование показало, что модульная организация БЛП предоставляет широкие возможности тестирования различных теорий, моделей и алгоритмов анализа без ущерба для работы остальных модулей, при условии соблюдения принципов, описанных в классификаторах свойств языка для получения сопоставимых результатов.

Предложенный БЛП имеет особую ценность для теоретических и практических лингвистических исследований, так как, благодаря возможности представления содержания текстовых документов одновременно на нескольких уровнях языковой системы, БЛП может использоваться для создания морфологических и лексико-грамматических корпусов текстов с целью разработки и тестирования вероятностно-статистических моделей, определения авторства работ, лексикологических и лексикографических исследований, создания конкордансеров, аннотаторов и других вспомогательных программ для ручной и полуавтоматической работы с текстами.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Выявлены особенности японского языка как объекта моделирования, наиболее существенно влияющие на эффективность решения задач его автоматического анализа. Впервые предложена концепция базового ЛП японского языка, которая в отличие от существующих ЛП является универсальной по отношению к решаемым задачам и ориентированной на их промышленный характер. Построена функциональная схема БЛП, согласно которой основные компоненты его ЛБЗ организованы в виде отдельных модулей, соответствующих функциональным уровням анализа текста [1; 5].
2. Разработан формат базового словаря японского языка, ориентированный на использование МЛЕ в качестве его словарных статей, а также их фонетической, графической и смысловой эквивалентности и словоизменительных свойств лексических единиц, которые в совокупности могут быть описаны на языке расширенных регулярных выражений, что существенно повышает производительность процесса построения БС и его обработки в составе БЛП [9].
3. Построена технология аннотирования базового корпуса текстов, основанная на сформулированных принципах определения границ МЛЕ и их ЛГК, а также на итеративном характере процесса, позволяющем постепенно переходить от его ручных процедур к автоматическим. Показано, что предложенный объём БКТ является достаточным для построения качественных вероятностных алгоритмов лексико-грамматического анализа текстов японского языка [6].
4. Разработано новое в совокупности лингвистическое обеспечение БЛП японского языка, включая классификаторы его свойств на лексико-грамматическом, синтаксическом и семантико-синтаксическом уровнях, базовый корпус текстов, базовый словарь и множества лингвистических правил нормализации текста, распознавания границ предложений, определения ЛГК неизвестных по отношению к базовому словарю МЛЕ, корректировки ЛГК МЛЕ, синтаксического и семантико-синтаксического анализа текста. Дана формализация лингвистических правил на языке расширенных регулярных выражений и построены принципиальные схемы их обработки на соответствующих этапах анализа текста [2; 3; 4; 5; 7; 9].
5. Разработаны и описаны в виде блок-схем алгоритмы нормализации текста и его графемного анализа, алгоритмы морфологического анализа текста и последующей корректировки его результатов, алгоритмы синтаксического и семантико-синтаксического анализа текстов. С

целью получения качественных характеристик разработанного БЛП японского языка построена технология тестирования всех его функциональных модулей, основанная на использовании эталонных корпусов текстов [5; 8].

6. Разработан прототип промышленного БЛП японского языка, осуществлено его внедрение и определены качественные характеристики, которые являются сопоставимыми, а в ряде случаев лучшими по сравнению с известными системами, решающими отдельные задачи автоматического анализа текстов японского языка. Совокупные же показатели качества построенного БЛП являются лучшими для лингвистических процессоров японского языка промышленного типа [5].

Рекомендации по практическому использованию результатов

Разработанный базовый лингвистический процессор, а также технологии и лингвистические ресурсы рекомендуются к использованию при создании лингвистических процессоров систем автоматической обработки текстовых документов (автоматического реферирования и аннотирования, машинного перевода, экспертных и вопросно-ответных систем, информационного поиска, и т.д.), а также в учебном процессе в высших учебных заведениях, осуществляющих подготовку специалистов в области интеллектуальных информационных систем и компьютерной лингвистики.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в научных журналах

1. Кравченко, С.Ю. Особенности японского языка как объекта моделирования с целью автоматического анализа текста / С.Ю. Кравченко // Вестник МГЛУ. Сер. 1, Филология. – 2012. – № 2 (57). – С. 140–147.
2. Кравченко, С.Ю. Лексико-грамматический классификатор японского языка / С.Ю. Кравченко // Вестник МГЛУ. Сер. 1, Филология. – 2012. – № 3 (58). – С. 108–116.
3. Кравченко, С.Ю. Синтаксический и семантико-синтаксический классификаторы японского языка / С.Ю. Кравченко // Вестник МГЛУ. Сер. 1, Филология. – 2012. – № 5 (60). – С. 117–127.
4. Кравченко, С.Ю. Особенности организации лингвистических правил в базе знаний системы автоматического анализа текстов японского языка / С.Ю. Кравченко // Вестник МГЛУ. Сер. 1, Филология. – 2013. – № 6 (67). – С. 134–141.
5. Кравченко, С.Ю. Базовый лингвистический процессор японского языка / С.Ю. Кравченко // Вестник МГЛУ. Сер. 1, Филология. – 2014. – № 1 (68). – С. 130–141.

Статьи в сборниках научных трудов

6. Кравченко, С.Ю. Особенности создания лексико-грамматического корпуса текстов научно-технической тематики для японского языка / С.Ю. Кравченко // Исследования молодых учёных: сб. статей аспирантов. – Минск, МГЛУ 2012. – С. 150–155.

Статьи в материалах научных конференций

7. Кравченко, С.Ю. Лексико-грамматический классификатор свойств японского языка / С.Ю. Кравченко // Материалы Междунар. конгресса по информатике: информационные системы и технологии (CSIST’2011). – Минск, 31 окт. –3 нояб. 2011. – Ч. 1. – С. 212–216.
8. Krauchanka S. Memory-Efficient Katakana Compound Segmentation Using Conditional Random Fields / S. Krauchanka, A. Artsymenia // In proc. of intl.

conf. on Computational Linguistics 2012. – Mumbai, 2012. – Posters Vol.3. – P. 1131–1139.

9. Кравченко, С.Ю. Базовый словарь системы автоматического анализа текстов японского языка / С.Ю. Кравченко // Материалы Междунар. конгресса по информатике: информационные системы и технологии (CSIST’2013). – Минск, 4–7 нояб. 2013. – Минск, 2013. – С. 148–153.

РЕЗЮМЕ

Кравченко Сергей Юрьевич

СИСТЕМА БАЗОВОГО АВТОМАТИЧЕСКОГО ЛИНГВИСТИЧЕСКОГО АНАЛИЗА ТЕКСТОВ ЯПОНСКОГО ЯЗЫКА

Ключевые слова: базовый лингвистический процессор, японский язык, автоматическая обработка текстов, автоматический морфологический анализ, автоматический синтаксический анализ, лексико-грамматический корпус текстов, лингвистические правила анализа текстов.

Объектом исследования являются процессы автоматического лингвистического анализа текстов японского языка на различных уровнях его глубины.

Предметом исследования являются методы, алгоритмы и лингвистические ресурсы базового автоматического лингвистического анализа текстов японского языка.

Целью диссертационной работы является разработка лингвистического, алгоритмического и технологического обеспечения базового лингвистического процессора японского языка и его реализация в виде промышленного прототипа.

В работе предложена концепция базового лингвистического процессора японского языка и его функциональная схема, разработан формат базового словаря японского языка, построена технология аннотирования базового корпуса текстов, позволяющая постепенно переходить от ручных процедур к автоматическим. Разработано новое в совокупности лингвистическое обеспечение базового лингвистического процессора японского языка, включая классификаторы его свойств на лексико-грамматическом, синтаксическом и семантико-синтаксическом уровнях, базовый корпус текстов, базовый словарь и множества лингвистических правил анализа текста. Дана формализация лингвистических правил на языке расширенных регулярных выражений и построены принципиальные схемы их работы на соответствующих этапах анализа. С целью получения качественных характеристик разработанного базового лингвистического процессора японского языка построена технология тестирования всех его функциональных модулей.

Разработан прототип промышленного базового лингвистического процессора японского языка и определены его качественные характеристики, которые являются сопоставимыми, а в ряде случаев лучшими по сравнению с известными системами, решающими отдельные задачи автоматического анализа текстов японского языка. Прототип внедрён в составе известной системы автоматизации инженерии и управления знаниями IHS Goldfire.

РЭЗЮМЭ

Краўчанка Сяргей Юр'евіч

СІСТЭМА БАЗАВАГА АЎТАМАТЫЧНАГА ЛІНГВІСТЫЧНАГА АНАЛІЗУ ТЭКСТАЎ ЯПОНСКАЙ МОВЫ

Ключавыя словы: базавы лінгвістычны працесар, японская мова, аўтаматычная апрацоўка тэкстаў, аўтаматычны марфалагічны аналіз, аўтаматычны сінтаксічны аналіз, лексіка-грамматычны корпус тэкстаў, лінгвістычныя правілы аналізу тэкстаў.

Аб'ектам даследавання з'яўляюцца працэсы аўтаматычнага лінгвістычнага аналізу тэкстаў на японскай мове на розных узроўнях яе глыбіні.

Прадметам даследавання з'яўляюцца метады, алгарытмы и лінгвістычныя рэсурсы базавага аўтаматычнага лінгвістычнага аналізу тэкстаў на японскай мове.

Мэтай дысертацыйнай работы з'яўляецца распрацоўка лінгвістычнага, алгарытмічнага і тэхналагічнага забеспячэння базавага лінгвістычнага працэсара японскай мовы і яго рэалізацыя ў выглядзе прамысловага прататыпа.

У рабоце прапанавана канцэпыя базавага лінгвістычнага працэсара японскай мовы і яго функцыянальная схема, распрацаваны фармат базавага слоўніка японскай мовы, пабудавана тэхналогія анатавання базавага корпуса тэкстаў, якая дазваляе паступова пераходзіць ад ручных працэдур да аўтаматычных. Распрацавана новае ў сукупнасці лінгвістычнае забеспячэнне базавага лінгвістычнага працэсара японскай мовы, у тым ліку класіфікатары яе ўласцівасцей на лексіка-грамматычным, сінтаксічным і семантыка-сінтаксічным узроўнях, базавы корпус тэкстаў, базавы слоўнік і мноства лінгвістычных правіл аналізу тэксту. Дадзена фармалізацыя лінгвістычных правіл на мове пашыраных рэгулярных выразаў і пабудаваны прынцыповыя схемы іх працы на адпаведных этапах аналізу. З мэтай атрымання якасных характарыстык распрацаванага базавага лінгвістычнага працэсара японскай мовы пабудавана тэхналогія тэставання ўсіх яго функцыянальных модуляў.

Распрацаваны прататып прамысловага базавага лінгвістычнага працэсара японскай мовы і вызначаны яго якасныя характарыстыкі, якія з’яўляюцца супастаўляльнымі, а ў шэрагу выпадкаў – праўзыходнымі ў параўнанні з вядомымі сістэмамі, што вырашаюць асобныя задачы аўтаматычнага аналізу тэкстаў на японскай мове. Прататып быў ўкаранёны ў складзе вядомай сістэмы аўтаматызацыі інжынерыі і кіравання ведамі IHS Goldfire.

SUMMARY

Krauchanka Siarhei

SYSTEM OF BASIC AUTOMATIC LINGUISTIC ANALYSIS OF JAPANESE TEXTS

Keywords: basic linguistic processor, Japanese language, automatic text processing, automatic morphological analysis, automatic syntactic analysis, part-of-speech text corpus, linguistic rules of text analysis.

Object of the research: processes of automatic linguistic analysis of Japanese texts on different language levels.

Subject of the research: methods, algorithms and linguistic resources of basic linguistic analysis of Japanese texts.

The purpose of the thesis is to develop linguistic, algorithmic and technological resources of the basic linguistic processor of the Japanese language, and to implement it as an industrial prototype.

The thesis proposes the concept and the functional scheme of a basic linguistic processor of the Japanese language, the format of basic dictionary of the Japanese language, and the technology, which allows gradual transition from manual to automatic annotation of a part-of-speech corpus. The thesis results in development of new linguistic resources for the basic linguistic processor of the Japanese language, such as: language property classifiers on lexico-grammatical, syntactic and semantico-syntactic levels; base corpus; base dictionary; the set of linguistic rules of text analysis. The format of linguistic rules is based on the formal language of word-based regular expressions. In order to evaluate the quality of the basic linguistic processor the technology of evaluation is described for each of its modules.

The prototype of the basic linguistic processor was developed and evaluated. The acquired quality metrics are competitive to, and in a number of cases, are higher than those of the existing systems, which are used for specific tasks of Japanese text processing. The prototype was implemented as a part of the well-known system of automatic knowledge engineering and management – IHS Goldfire.

Научное издание

КРАВЧЕНКО
Сергей Юрьевич

СИСТЕМА БАЗОВОГО АВТОМАТИЧЕСКОГО ЛИНГВИСТИЧЕСКОГО АНАЛИЗА ТЕКСТОВ ЯПОНСКОГО ЯЗЫКА

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата филологических наук

по специальности 10.02.21 – прикладная и математическая лингвистика

Ответственный за выпуск *С.Ю. Кравченко*

Подписано в печать 12.11.2014 г. Формат 60x84 1/16. Бумага офсетная. Гарнитура Times New Roman. Ризография. Усл. печ. л. 1,63. Уч.-изд. л. 1,31. Тираж 100 экз. Заказ 64 .

Издатель и полиграфическое исполнение: учреждение образования «Минский государственный лингвистический университет». Свидетельство о государственной регистрации издателя, изготовителя, распространителя печатных изданий от 02.06.2014 г. № 1/337. ЛП № 02330/458 от 23.01.2014 г.
Адрес: ул. Захарова, 21, 220034, г. Минск.