УДК 004.912

Комплексный подход к исследованию лексических характеристик текста

Е. А. Сидорова[1]

В работе предлагается подход и рассматривается программное обеспечение для многоцелевого исследования лексических характеристик текста. Данная работа лежит на стыке корпусной лингвистики и лексикографических исследований. Основой проводимых исследований является корпус текста и создаваемый на его основе проблемно-ориентированный словарь. Необходимое программное обеспечение поддержки исследователя включает интерфейсы для разработки словарей, построения системы признаков, разметки терминов, а также средства автоматической генерации лексического наполнения словаря по текстам, поиска контекстов терминов, накопление статистической информации и др. При извлечении терминов осуществляется морфологический анализ текста и построение словосочетаний на основе правил согласования грамматических характеристик слов. Для исследования контекстов употребления терминов предоставляются средства построения конкордансов, что позволяет конечному пользователю пронаблюдать грамматические, семантические, стилистические и проблемно-ориентированные особенности терминов и осуществить их разметку.

Ключевые слова: термин, предметный словарь, корпус текстов, конкорданс.

1. Введение

Текст как источник и средство передачи информации нуждается во всестороннем исследовании, необходимом как для оценки качества изложенного, так и при автоматической обработке и поддержке языковых поисковых сервисов. Изучение языковых явлений и моделирование процессов понимания текста на разных языковых уровнях является фокусом современных исследований в компьютерной лингвистике. Для решения данных задач, как правило, используют разнообразные знания в формализованном виде, такие как тезаурусы (например, WordNet), толково-комбинаторные словари, аннотированные корпуса текстов (например, Национальный корпус русского языка, `www.ruscorpora.ru`) и т.п. Тезаурус как инструмент описания предметной лексики позволяет характеризовать термин и его связи с точки зрения особенностей употребления в данной предметной области [1]. Другим вариантом исследования языковых явлений является использование корпусов текстов. Корпус является источником и инструментом многоаспектных лексикографических работ [2], позволяя на основе разметки автоматизировать создание и начальное наполнение словарей.

Анализ литературы [3–5] показывает, что при извлечении терминологии из большого массива текстов используются подходы, объединяющие лингвистические и статистические методы. Результатом такого рода исследований может стать аннотированный корпус текстов, который позволит в дальнейшем осуществлять частотный анализ лексики, создавать конкор-

[1] Исследование выполнено при поддержке РФФИ в рамках проектов №17-07-01600 и № 18-00-01376 (18-00-00889).

дансы по различным основаниям и строить электронные словари. Использование специализированных методов позволит автоматизировать работу экспертов по исследованию формальных структур и дискурса в целом и построению лингвистических моделей на основе размеченного корпуса текста.

Данная работа посвящена описанию методики и инструментов, обеспечивающих исследования лексических характеристик текста на основе корпусов текстов. Совокупность предложенных инструментов позволяет сформировать среду для создания проблемно-ориентированных словарей и предоставить конечному пользователю различные возможности исследования языковых явлений.

2. Среда поддержки исследования текста

Разработка моделей и создание качественных ресурсов требует кропотливого ручного труда, поддерживаемого программным инструментарием. Программная среда должна предоставлять специалистам различные рабочие инструменты для формирования необходимых баз знаний и проведения корпусных исследований.

В рамках данной работы были сформулированы функциональные требования, которым среда должна удовлетворять:

a) обеспечивать автоматическое наполнение словарей на базе корпусов текстов;

b) предоставлять возможность настраивать и приписывать различные характеристики терминам словаря;

c) осуществлять лексический анализ текста – сегментацию и извлечение из текста заданных в словаре терминов;

d) обеспечивать накопление данных о статистико-комбинаторных свойствах обнаруженных в тексте языковых явлений;

e) осуществлять построение конкордансов терминов и их визуализацию.

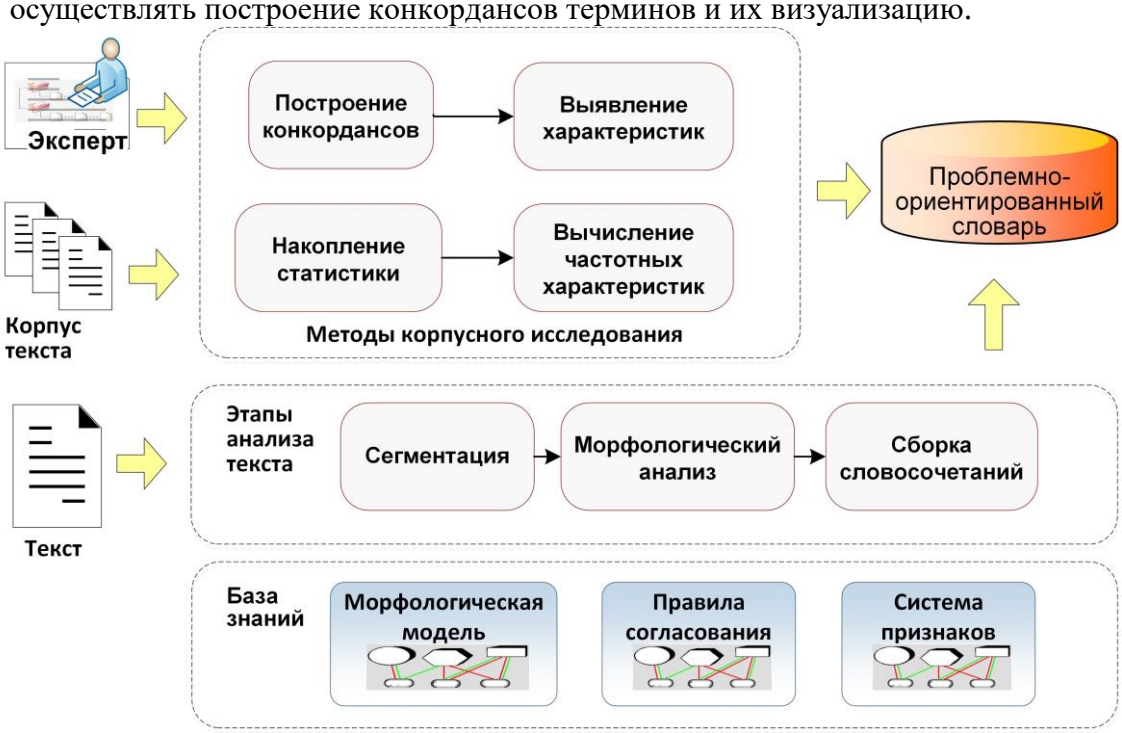

Рис. 1. Среда исследования лексических характеристик текста

Созданная система включает следующие базовые инструменты для проведения исследований (рис. 1): интерфейс разработки словаря и построения системы признаков, средства автоматической генерации лексического наполнения словаря по корпусу текстов и вычисления

количественных характеристик найденных терминов, инструменты построения конкордансов для исследования контекстов лексических единиц.

3. Модель представления знаний

Рассматриваемая лексикографическая модель знаний включает три основных компонента. Словарь задает лексическую модель рассматриваемого подъязыка, определяемого проблемной областью. Грамматика обеспечивает поиск и извлечение лексических единиц из текстов. Система прагматически ориентированных характеристик, задаваемая пользователем, поддерживает фиксацию наблюдений и ориентирована на дальнейшую поддержку автоматизированных методов обработки текста.

Базой исследований является представительный проблемно-ориентированный корпус текстов. Основным инструментом, обеспечивающим поддержку исследований, является поиск примеров употребления терминов словаря, построение множества контекстов (конкордансов), вычисление количественных характеристик встречаемости, совместной встречаемости, распределения и др.

3.1. Лексическая модель

В рамках нашего подхода в словарной статье хранится необходимая информация как для извлечения термина из текста, так и для поддержки последующих этапов анализа.

Проблемно-ориентированный словарь – это объем лексики, организованной по семантическому (тематическому / жанровому / и др.) принципу с отражением определенного набора базовых формальных отношений. Формально словарь описывается системой вида:

$D = \{W, P, M, G, S, F_w, F_p\}$,

где W – множество лексем, каждой лексеме сопоставлена информация обо всей совокупности ее форм;

P – множество многословных терминов, описываемых парой <N-грамма, тип структуры>, где N-грамма задает последовательность лексем, а тип структуры определяет вершину и правила согласования элементов N-граммы;

M – морфологическая модель языка, включающая описание морфологических классов и признаков;

G – множество правил согласования для извлечения многословных терминов;

S – проблемно-ориентированная система признаков для разметки терминов;

$F_w = W \to 2^{M \times S}$, $F_p = P \to 2^{G \times S}$ – функции, сопоставляющие терминам наборы признаков.

Особенностью морфологического представления в рамках системы является возможность ее изменения в зависимости от задачи, решаемой с помощью словаря. Пользователь может сформировать собственный набор признаков, классов и обеспечить их интеграцию посредством правил сопоставления с базовым морфологическим представлением. Потребность в изменении классов возникает достаточно редко, например, когда используются дополнительные специализированные словари терминов (словари имен, географических названий) или необходимо включить в словарь слова другого языка.

Морфологический класс определяется частью речи, набором лексических признаков (например, одушевленность или род у существительных) и типом парадигмы.

Морфологический атрибут – это пара $<t^m, V^m>$, где t^m задает тип признака (имя), а V^m – множество значений (например, <Род, {муж, жен, ср}>). Атрибуты в рамках каждого класса делятся на словообразующие или лексические, присущие всем формам лексем данного класса, и словоизменительные, различающие формы одной лексемы.

Для описания словоизменения используется понятие парадигмы – совокупности всех словоформ, относящихся к одной лексеме и имеющих разные грамматические значения. Тип парадигмы задает упорядоченный набор изменяемых атрибутов лексем данного морфологи-

ческого класса (например, для обычного прилагательного это тройка <падеж, число, род>) и функции для вычисления значений признаков для заданной словоформы (точнее, ее флексии) и наоборот. Таким образом, каждой лексеме сопоставляется морфологический класс и парадигма из таблицы парадигм, а для каждой парадигмы можно вычислить значения морфологических признаков относительно типа парадигмы, определяемого морфологическим классом.

Другой важной особенностью системы является поддержка многословных терминов – словосочетаний, сформированных по правилам, реализующих поверхностно-синтаксический анализ. Большинство многословных терминов включают от двух до четырех слов и формируются с помощью правил вида П+С (*аналоговый датчик*) – согласование существительного и прилагательного, С+Срд (*автор учебника*), П+П+С (*новая информационная технология*), С+Прд+Срд (*обработка естественного языка*) и т.п. Встречаются также термины с более сложной структуры, например, с зависимыми предложными группами С+Предл+С (*резервуар с жидкостью*), С+Предл+С+С (*поиск в пространстве состояний*) и т.д.

3.2. Система проблемно-ориентированных признаков

В зависимости от решаемой проблемы термины словаря могут снабжаться различными типами признаков: статистическими (для решения задач классификации), жанровыми (для жанрового анализа текста), семантическими (для семантического анализа), формальными (для выявления маркеров определенных структур) и др.

Статистические признаки накапливают информацию о частоте появления термина в обрабатываемых текстах. Для решения задач классификации текста необходимо, чтобы пользователь задал систему связанных между собой тем, а также наличие обучающего корпуса текстов, т.е. корпуса, размеченного заданными признаками. В словаре для каждого термина хранится: встречаемость в обучающей выборке (абсолютная частота) и количество текстов, в которых встречался термин (текстовая частота), список тем, в которых встретился термин, частота и текстовая частота по каждой теме из списка. Часть параметров (относительная частота, TF*IDF, вес) вычисляется динамически.

Система признаков, используемая для разметки терминов словаря, формируется пользователем и зависит от решаемой задачи, т.е. является проблемно-ориентированной. Для кодирования различной информации о слове (семантической, жанровой, стилистической и др.) предусмотрены следующие возможности.

Класс. Термин может быть отнесен к определенному признаку-классу. Иерархия классов позволяет отнести термин к определенному уровню иерархии, более общему или конкретному с наследованием свойств общего класса.

Атрибут. Для представления лексического значения термина используются атрибуты. Совокупность семантических значений атрибутов, приписанных слову, в определенной мере моделирует компонентную семантическую структуру слова. Основные компоненты семантической структуры термина могут рассматриваться как тезаурусные дескрипторы.

Механизм фиксации значений. Набор признаков (классов, атрибутов и их значений) позволяет достаточно полно описать лексико-семантическое значение термина. Наличие нескольких альтернативных групп позволяет выразить неоднозначность термина.

Рассмотрим семантическую разметку многозначного термина *нефтедобыча* (рис. 2).

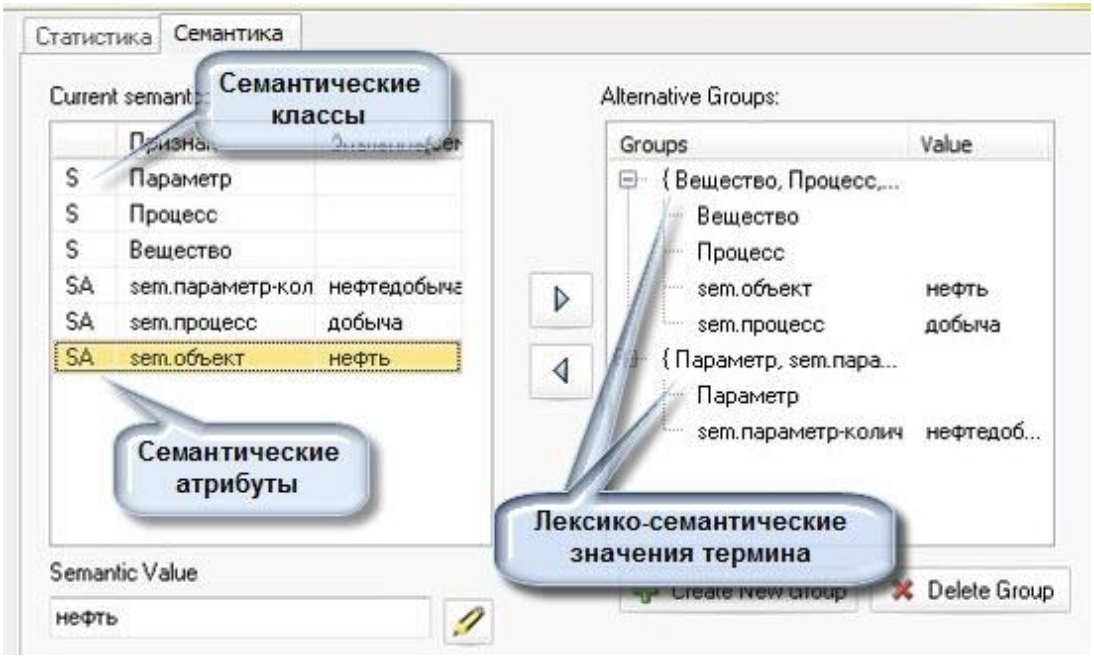

Рис. 2. Разметка термина *нефтедобыча*

При разметке термина вначале фиксируются все возможные признаки термина, а затем осуществляется фиксация значений. Такой механизм позволяет задавать как альтернативные, так и синкретически выраженные значения терминов. Так, в примере, приведенном на рис. 2, термину приписано два лексико-семантических значения: а) процесс *Добыча* над *Веществом* и б) параметр *нефтедобыча*, значение которого может выражаться числовым значением. Дополнительно значения снабжены семантическими атрибутами, позволяющими задать канонические названия признаков.

4. Методы поддержки работы с корпусом текстов

Созданная среда включает словарные компоненты и обработчики, которые обеспечивают, с одной стороны, автоматизацию создания, наполнения, редактирования словарей, с другой – применение созданных словарей для лексического анализа текста и последующую работу со словарной информацией найденных в тексте терминов. Важной частью поддержки пользователя является набор поисковых средств, таких как сортировка и фильтрация терминов по различным наборам параметров, визуализация покрытия текста терминами словаря, построение конкордансов терминов с расширяемым окном контекста и др.

4.1. Обучение словаря на основе корпуса текстов

Процесс извлечения терминологии включает такие этапы, как: a) графематический анализ, обеспечивающий токенизацию и выделение нетекстовых элементов и иноязычных вкраплений, b) лексико-морфологический анализ (лемматизация, определение лексико-грамматических признаков, представление парадигмы, нормализация), c) выделение терминоподобных словосочетаний (идентификация на основе предопределенных грамматических моделей и нормализация), d) обновление статистики найденных терминов.

Для создания словарей используются следующие модули.

Морфологический анализ осуществляется на базе модуля Диалинг (`www.aot.ru`), который содержит универсальный словарь русского языка и обеспечивает поиск слова в словаре, определение его грамматических признаков и нормальной формы. Также поддерживается функция предсказания [6], которая по незнакомому слову формирует гипотезы (как правило, около трех вариантов) о его части речи, нормальной форме и других признаках.

Модуль сборки словокомплексов – извлекает из текста словосочетания по фиксированному набору грамматических правил. Основной задачей модуля является выявление наиболее важных терминообразующих синтаксических групп, большинство из которых представляют собой именные группы либо строятся на их основе.

Результатом обработки корпуса текста с помощью данных модулей является лексическое наполнение словаря с одновременным сбором статистики встречаемости. Если корпус размечен признаками, то соответствующие термины снабжаются данными признаками и относительно признаков также ведется статистика.

Таким образом, среда обеспечивает автоматизацию начального наполнения словарей, на основе которого можно проводить дальнейшие исследования.

4.2. Исследование контекстов термина

Конкорданс – традиционный способ изучения корпуса текста. Он дает полный индекс терминов в ближайших и расширенных контекстах, что дает возможность исследователю проверить свою гипотезу относительно функций той или иной лексической единицы. Конкорданс создается для поиска и извлечения образцов целевых единиц. Таким образом, конкорданс осуществляет обратную связь словаря, словарных терминов с корпусом и обеспечивает своего рода лингвистическую разметку на морфологическом и поверхностно-синтаксическом уровне.

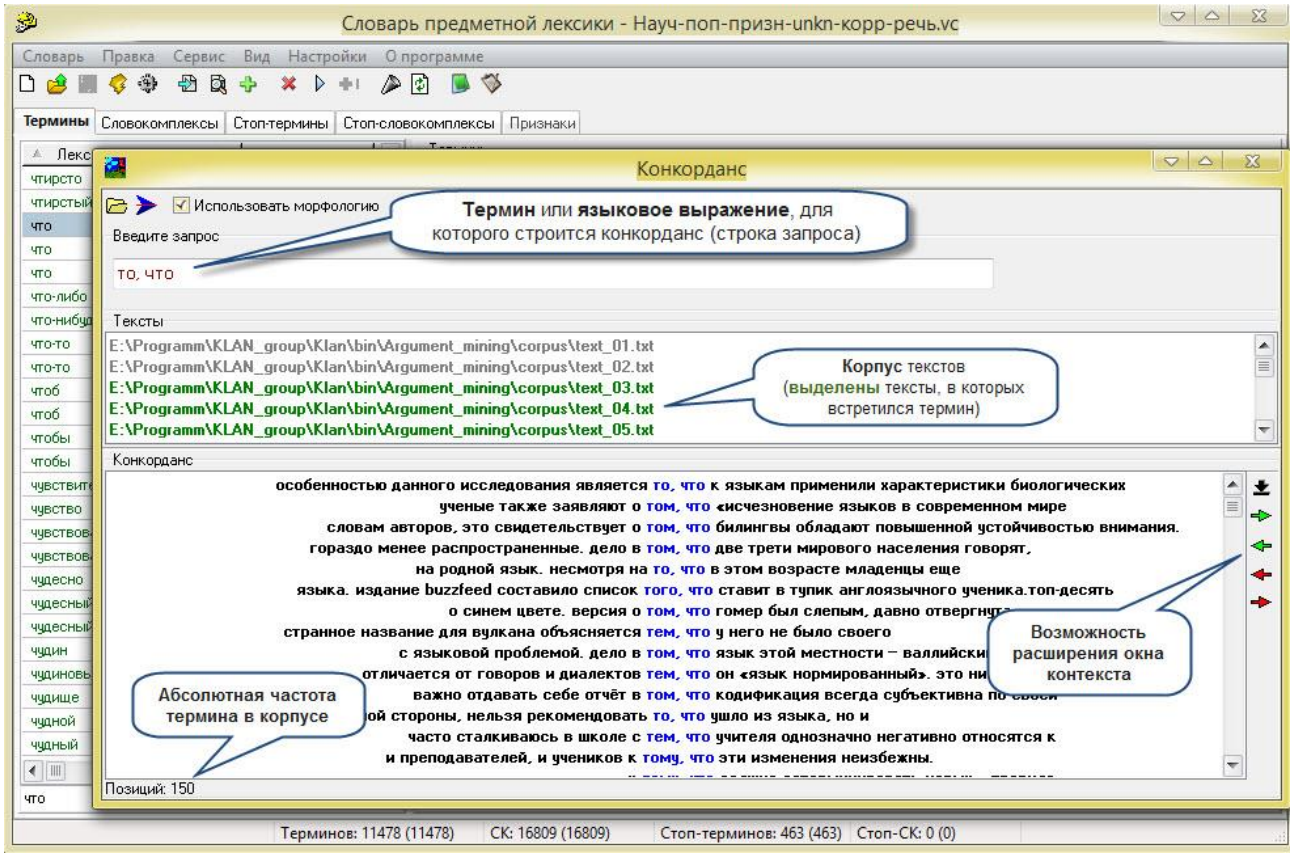

Рис. 3. Исследование контекстов термина с помощью конкорданса

Разработанный модуль конкорданса (рис. 3) работает с небольшими корпусами текстов и осуществляет построение конкордансов в реальном времени, позволяя пользователю уточнять и изменять запрос. При просмотре множества контекстов вхождения термина пользователь может самостоятельно определять длину просматриваемого фрагмента текста (поддерживается пословное расширение контекста, просмотр абзаца или всего сообщения). Так, в приведенном примере вначале был сформирован конкорданс для термина *что* (869 вхожде-

ний), а затем запрос был расширен до языковой конструкции *то, что* (150 вхождений). Более подробный анализ контекста позволил определить, что данная конструкция позволяет выделить аргументацию, представленную косвенной речью эксперта, с точностью ~ 65 %. А при включении в модель описания аргумента речевых и ментальных предикатов и терминов с семантикой человека-эксперта точность распознавания аргументации увеличилась до 86.5 %.

Таким образом, необходимо отметить, что конкорданс является не только средством исследования контекстов терминов, но и важным инструментом проверки гипотез относительно функционирования языка и создания качественных лингвистических ресурсов.

4.3. Методика исследования лексических характеристик текста

В процессе создания лингвистического ресурса можно выделить несколько содержательных этапов (рис. 4): сбор и разметка данных, первоначальное наполнение словаря на основе корпуса, формирование системы признаков для моделирования значений терминов, исследование терминов с помощью конкорданса, генерация гипотез и их экспертная оценка. Полученный ресурс можно использовать для автоматизации разметки новых данных.

Рис. 4. Жизненный цикл разработки лингвистических ресурсов

Процесс первоначального наполнения словаря осуществляет извлечение из массива текстов терминов-кандидатов на базе определенных лингвистических ограничений (морфологических и синтаксических моделей) с использованием технологии обучения по массиву текстов. Морфологический анализ текстов (на основе русского морфологического анализатора) позволяет извлечь однословную лексику. Учитывая, что многие термины специализированных предметных областей универсальному словарю-анализатору неизвестны, используются средства предсказания незнакомых слов. Поверхностный синтаксический анализ выделяет в текстовых сегментах контактные синтаксически связанные группы слов (словокомплексы) на основе фиксированного набора правил.

Система признаков, используемая для разметки текстов, зависит от предметной и проблемной области, её формирование является достаточно сложной научно-исследовательской проблемой. Так, для создания ресурсов, ориентированных на автоматическое извлечение информации или информационный поиск, система признаков может быть построена на основе онтологии предметной области и дополнена необходимыми признаками для решения стандартных лингвистических проблем. Еще одним широко используемым методом построения системы признаков является использование универсальных тезаурусов для фиксации верхнего уровня иерархии дескрипторов с дальнейшим расширением специализированными при-

знаками [1]. В задачах классификации текстов применяются методы для сужения пространства признаков.

Разметка терминов может быть сформирована автоматически при наличии размеченного корпуса текста. В противном случае разметка терминов осуществляется экспертом вручную. Работа эксперта поддерживается инструментальными средствами сортировки и фильтрации терминов по различным наборам признаков, поиска всех многословных терминов, содержащих слово, построения конкорданса терминов с учетом словоизменения. Механизм фиксации значений терминов позволяет достаточно полно описать его особенности.

Формирование более сложных лингвистических моделей и их проверка относительно лексической составляющей позволяет добиться необходимого качества создаваемого ресурса.

5. Заключение

Данная работа посвящена описанию подходов и методов разработки лексикографических ресурсов и проведения корпусных исследований для обеспечения полноты и достоверности разрабатываемых моделей. В фокусе внимания создаваемого инструментария находятся лексические единицы, выступающие в качестве маркеров и индикаторов объектов более высокого уровня (семантического, прагматического, тематического, структурно-жанрового, логико-аргументативного и др.).

Отличительными особенностями рассматриваемого программного комплекса являются возможность его многоцелевого использования при решении различных задач анализа текста, интеграция различных возможностей для проведения языковых исследований, обеспечение настройки словарей на проблемную область пользователя, поддержка мультиязычности.

Развитие предлагаемого подхода заключается в апробации и внедрении методов автоматического построения моделей, ориентированных на извлечении прагматической информации из текстов: описание объектов и их параметров, моделей для извлечения отношений (ситуаций), жанровых особенностей текста, риторической аргументации и т.п.

Литература

1. *Лукашевич Н. В.* Тезаурусы в задачах информационного поиска. М.: МГУ, 2011. 495 с.
2. *Sinclair J.* Corpus, Concordance, Collocation. Edited by Ronald Carter. Oxford: Oxford University Press, 1991, XVIII, 179. 200 p.
3. *Захаров В. П., Хохлова М. В.* Автоматическое выявление терминологических словосочетаний // Структурная и прикладная лингвистика. 2014. Вып. 10. С. 182–200.
4. *Bolshakova E., Loukachevitch N., Nokel M.* Topic Models Can Improve Domain Term Extraction // International conference on Information Retrieval (ECIR-13), Springer Verlag, 2013. LNCS-7814. P. 684–687.
5. *Митрофанова О. А., Захаров В. П.* Автоматизированный анализ терминологии в русскоязычном корпусе текстов // Компьютерная лингвистика и интеллектуальные технологии: тр. межд. конференции «Диалог-2009». С. 321–328.
6. *Сокирко А. В.* Морфологические модули на сайте www.aot.ru // Компьютерная лингвистика и интеллектуальные технологии: тр. межд. конференции Диалог-2004. С. 559–564.

Статья поступила в редакцию 19.07.2019;
переработанный вариант – 05.08.2019.

Сидорова Елена Анатольевна

к.ф.-м.н., с.н.с., лаборатория искусственного интеллекта, Институт систем информатики им. А. П. Ершова СО РАН (630090, Новосибирск, просп. Академика Лаврентьева, 6), тел. (383) 3-307-991, e-mail: lsidorova@iis.nsk.su.

The integrated approach to text lexical characteristics study

E. Sidorova

The integrated approach and software environment for multi-aspect study of the text lexical characteristics are considered. This work is at the junction of corpus linguistics and lexicographical research. The basis of the research is the corpus of text and the problem-oriented dictionary. The proposed environment for supporting the researcher provides tools and interfaces for developing vocabularies and a system of domain features, terms markup, automatic generation of lexical content and accumulation of statistical information, etc. To extract terms the morphological analysis and the construction of phrases based on the rules of matching the grammatical characteristics of words are carried out. To study the contexts of the terms use, concordance construction tools are provided. Concordances allow the researcher to test his or her hypothesis about the functionality of a particular lexical unit. The considered environment allows to solve various text analysis tasks because it integrates various tools for conducting language research and supports customization of vocabularies to a problem area.

Keywords: terminology, domain vocabulary, corpora, concordance.