ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Учебное пособие предназначено для студентов вузов, осваивающих образовательные программы бакалавров и магистров по направлению 230400 — «Информационные системы», а также 230100 — «Информатика и вычислительная техника». Материал учебных пособий ориентирован на дисциплины как базовой, так и вариативной частей профессионального цикла соответствующих Федеральных государственных образовательных стандартов.

Пособие состоит из двух частей. Обе части посвящены рассмотрению компьютерных технологий, методов и средств обработки текстовой информации, извлечению знаний из текстовых данных.

Часть 1, получившая название «Начало…», посвящена технологиям, которые использовались на начальном периоде развития методов обработки текстовой информации с использованием ЭВМ. В основном это методы применения вычислительной техники в структурной, вычислительной и квантитативной лингвистике, которые использовались в информационно-справочных и библиотечных системах для поиска документов, для создания первых информационно-справочных тезаурусов. Она базируется на исследованиях и разработках последней четверти XX века, снабжена библиографическим справочником по книжным публикациям этого периода (в основном отечественных, но также и зарубежных авторов), которые есть в библиотечных фондах высших учебных заведений.

Часть 2 названа «Text mining…». В ней рассматривается развитие в первом десятилетии XXI века методов обработки текстов последней четверти XX века. Она посвящена описанию современных методов обработки коллекций текстовых документов, информации размещенной в Интернет, а также в базах данных корпоративных информационных систем. В ней представлены методики и алгоритмы автоматизированного извлечения семантических отношений между понятиями в текстах для построения тезаурусов.

Книга «Лингвистическое обеспечение информационных систем. Часть 1. Компьютерная лингвистика. Начало (посл.четв. XX в.)» является учебным пособием по дисциплинам «Лингвистическое обеспечение информационных систем» и «Компьютерная лингвистика». Основной целью пособия является обеспечение необходимыми учебно-методическими материалами этих учебных дисциплин и способствование в приобретении и развитии слушателями следующих укрупненных компетенций: «Лингвистика как предметная область научных исследований и разработок», «Количественная спецификация естественно-языковых знаковых систем»

Основными образовательными результатами являются знания, умения, навыки и представления. В результате изучения материала дисциплины «Компьютерная лингвистика» студенты должны:

Знать: методы и решения в системах организации знаний; эмпирические законы ЕЯ описания; логико-статистические методы анализа ЕЯ описания предметных областей; технологию автоматизированной обработки текстовой информации.

Уметь: использовать технологию автоматизированной обработки текстовой информации для анализа ЕЯ описаний предметных областей.

Иметь навыки (владеть навыками): работы со специальными программными средствами автоматизированной обработки текстов.

Иметь представление: о системе языковедческих дисциплин; предмете, методах и моделях прикладной лингвистики; предмете компьютерной лингвистики.

Структура курса «Компьютерная лингвистика» включает следующие модули:

Модуль 1. Введение в компьютерную лингвистику.

Модуль 2. Исследование естественно-языковых (ЕЯ) ресурсов.

Модуль 3. Программные средства анализа текстов.

Основные формы учебных занятий — это лекции и практические занятия (лабораторные работы). Виды самостоятельной внеаудиторной работы: самостоятельная работа над материалом лекций и лабораторных работ — конспектирование научных статей по темам лекций, составление индивидуальных тематических словарей, решение задач; домашнее задание.

Данное учебное пособие состоит из пяти глав, справочного аппарата и приложений.

Первая глава «Лингвистика как предметная область научных исследований и разработок», может раскрываться в одной лекции. В главе рассматриваются два вопроса: 1) система языковедческих дисциплин и 2) компьютерная лингвистика.

В материалах по первому вопросу сначала приводятся определения объекта, предмета и содержания науки «Языкознание». Затем подробно рассматривается содержание прикладных лингвистических исследований, ее предмет, методы и модели.

В материалах по второму вопросу показывается междисциплинарный характер компьютерной лингвистики, как одной из основных прикладных лингвистических дисциплин. Далее анализируется инструментарий компьютерной лингвистики, а затем приводятся сведения об основных направлениях исследований и разработок.

В конце главы приводятся контрольные вопросы и задания.

В материалах главы использованы суждения и работы следующих ученых: В.М.Андрющенко, Н.Д.Арутюновой, Б.Ю.Городецкий, Ю.Н.Караулов, Ю.С.Степанова. Зна-

чительная часть материалов почерпнута из учебного пособия А.Н.Баранова «Введение в прикладную лингвистику».

Во второй главе «Исследование естественно-языковых ресурсов» рассмотрены основные теоретические вопросы и проблемы, связанные с формализацией знаний в современных информационных системах и их представлением в словарно-тезаурусной форме. Все рассматриваемые вопросы могут быть представлены в восьми лекциях, каждая из которых соответствует одному подразделу (параграфу).

В лекции 1 могут быть рассмотрены модели и методы представления и организации знаний, различные методы и решения в системах организации знаний и словарно-тезаурусное их представление: формализация и автоформализация знаний, лексикографическое (словарное) и логико-интуитивное описание, обобщение форм представления знаний и его предпосылки.

Лекции 2, 3 и 4 могут быть посвящены рассмотрению вопросов спецификации ЕЯ систем. Среди них: методы анализа ЕЯ описания предметных областей — модель «ранг-частота», законы Ципфа и Мандельброта, получение статистических распределений «ранг-частота»; логико-статистические методы извлечения знаний — дистрибутивно-статистический (ему следует посвятить целую лекцию), а также другие синтетические методы, среди которых частотно-семантический метод, компонентный анализ и различные интегрированные метрики.

Лекции 5 и 6 могут быть посвящены рассмотрению синтагматической и парадигматической моделям текста. Сначала следует формальное описание основных и производных синтагм, затем описываются синтагматические конструктивы и их статистический анализ, и в заключении парадигматические конструктивы. В лекции 5 можно проанализировать модель текста, а в лекции 6 представить технология автоматизированного построения словаря-тезауруса, в рамках которой рассматриваются: лингвистическая база данных, карта понятия и технология построения на их основе иерархической семантической сети.

В лекции 7 возможно привести пример исследования ЕЯ описания ПО и рассмотреть: построение ядра, генерального словника, семантической сети и карт понятий ЕЯ описания ПО.

Лекция 8 — заключительная. В ней можно привести описание словаря-тезауруса ПО «Информатика и вычислительная техника», который был получен с использованием рассмотренных в модуле методов.

В конце главы приводятся контрольные вопросы и задания.

В материалах этой главы пособия имеются ссылки на работы следующих ученых: К.Б. Бектаев, Р.Г.Пиотровский, Г.Г.Белоногов, Б.А.Кузнецов, Т.А.Гаврилова,

К.Р.Червинская, В.Е. Гмурман, Г.Р.Громов, Ю.Н.Караулов, И.А.Мельчук, В.И.Ракитин, В.Е.Первушин, А.Б.Соломоник, Ю.И.Шемакин, Ю.А. Шрейдер, А.Я,Шайкевич, В.А.Москович, Дж.Гласс, Дж.Стенли, Л.Закс, Б.Мандельброт, Дж.Сэлтон и др. Названы имена многих зарубежных и российских ученых.

В третьей главе «Программные средства анализа текстов» приведено описание программного комплекса «Система автоматизированного анализа естественно-языкового описания предметной области — Интерлекс 2.0». Данный комплекс предназначен для автоматизации исследований различных текстов и является основным инструментом, овладение которым позволит слушателям сформировать умения обработки естественно-языковой информации.

Содержание главы раскрывается в восьми лекциях. На первой лекции рассматриваются общая характеристика системы, ее архитектура и правила установки на компьютер. Далее описывается интерфейс: структура, системные функции, режимы. Дальнейшее изложение материала представляет собой рассмотрение работы программного комплекса в различных режимах: «Словник», «Словоформы», «Словарь», «Дерево», «Сеть». Содержание заключительной лекции — это рассмотрение структуры лингвистической базы данных, которая образуется в результате анализа текстового материала.

В конце главы приводятся контрольные вопросы и задания.

Материал главы базируется на книге — *Филиппович Ю.Н., Прохоров А.В.* Семантика информационных технологий: опыты словарно-тезаурусного описания — в ней наиболее полно описан программный комплек Интерлекс.

Четвертая глава «Библиотека «Начала компьютерной лингвистики»» содержит аннотированный список литературы по дисциплине «Компьютерная лингвистика» и включает аннотации, оглавление и др. материалы 31 книжного издания последней четверти XX века ведущих отечественных и зарубежных ученых осуществленных в нашей стране.

Пятая глава «Практикум» содержит две лабораторные работы на основе программного комплекса «Интерлекс 2.0», домашнее задание «Анализ информационного ресурса» и требования к самостоятельной работе студентов над материалом лекций.

Справочный аппарат пособия представлен поглавным списком литературы, на которую имеются ссылки в тексте пособия, списком основных сокращений, именным указателем и словарем терминов (глоссарием).

В словаре представлены основные термины и ключевые понятия дисциплины упорядоченные по темам и алфавиту. Содержание большинства словарных статей заимствовано из печатных или электронных лексикографических изданий (словарей, энциклопедий, справочников). В некоторых случаях для одного термина или понятия приведены несколько различ-

ных словарных статей, их структура определяется источником, сведения о котором приводятся в соответствующем поле. Для части терминов и понятий приведены авторские дефиниции и/или дескрипции из текста лекций.

В приложении к пособию представлены примерная программа дисциплины «Компьютерная лингвистика», модель компетенций, методические указания для студентов по освоению разделов пособия и тестовая система.

Благодарности:

Рецензентам В.Н. Агееву и В.Н.Полякову за полезные советы по содержанию рукописи пособия.

Семье, особенно мое дочери к.т.н., доценту Филиппович Анне Юрьевне. Она взяла на себя труд по организации издания пособия, оказала помощь в его редактировании.