Федеральное государственное автономное образовательное учреждение высшего образования
"Национальный исследовательский университет "Высшая школа экономики"

Факультет Компьютерных наук
Департамент анализа данных и искусственного интеллекта

Рабочая программа дисциплины «Компьютерная лингвистика и анализ текста»

для образовательной программы «Науки о данных»
направления подготовки 01.04.02. Прикладная математика и информатика
уровень - магистр

Разработчик программы
Большакова Е.И., кандидат физ.-мат. наук, доцент, eibolshakova@hse.ru

Одобрена на заседании департамента анализа данных и искусственного интеллекта
«___»_____ 2015 г.
Руководитель департамента С.О. Кузнецов _____

Рекомендована Академическим советом образовательной программы
«___»_____ 2015 г., № протокола_____

Утверждена «___»_____ 2015 г.
Академический руководитель образовательной программы
А.С. Конушин _____

Москва, 2015

1. Аннотация

Дисциплина «Компьютерная лингвистика и анализ текста» охватывает изучение различных моделей автоматической обработки текста на естественном языке, применяемых в современных информационных системах и затрагивающих несколько языковых уровней обрабатываемого текста, включая уровни морфологии, синтаксиса, дискурса и семантики. Изучаются также виды лингвистических ресурсов, используемых при обработке текстов, и методы их создания. Рассматриваются прикладные задачи, требующие многоуровневого анализа и синтеза текста (такие как машинный перевод, генерация текста, извлечение информации и знаний из текста).

2. Область применения и нормативные ссылки

Настоящая программа устанавливает минимальные требования к знаниям и умениям студента и определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих данную дисциплину, и студентов первого года обучения по направлению 01.04.02 «Прикладная математика и информатика», обучающихся по магистерской программе «Науки о данных» и выбравших для изучения данную дисциплину.

Программа разработана в соответствии с:

- Образовательным стандартом ВПО ГОБУ НИУ ВШЭ;
- Образовательной программой «Науки о данных» подготовки магистра направления 01.04.02 «Прикладная математика и информатика»;
- Рабочим учебным планом подготовки магистра по направлению 01.04.02, утвержденным в 2015 г.

3. Цели освоения дисциплины

Основная задача курса – изучение формальных моделей автоматической обработки текстов на естественном языке (ЕЯ). Эти модели применяются в различных типах прикладных информационных систем, и их освоение необходимо для подготовки специалистов в области наук о данных.

4. Компетенции, формируемые в результате освоения дисциплины

В результате изучения дисциплины студенты должны:

- Знать основные уровни анализа и синтеза текста на ЕЯ, существующие модели статистического, морфологического и синтаксического анализа текстов и их применение в типичных прикладных программных системах обработки текстов;
- Понимать существенные отличия естественных языков от искусственных и особенности компьютерных моделей естественного языка;
- Знать принципы построения различных лингвистических ресурсов, включая корпуса текстов, терминологические словари, тезаурусы, онтологии;
- Уметь применять существующие инструментальные средства и лингвистические ресурсы для разработки прикладных систем обработки текстов на естественном языке.

В результате изучения дисциплины студент осваивает и развивает следующие компетенции:

Компетенция	Код по ФГОС/ НИУ	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции
Способность строить и решать математические модели в соответствии с направлением подготовки и специализацией	ИК-М7.2 пми	Студент владеет основными моделями компьютерной лингвистики, используемыми в прикладных информационных системах	Изучение на лекциях методов автоматической обработки текстов и принципов построения лингвистических ресурсов, а также проведение на практических занятиях их сравнительного анализа на примерах работы известных прикладных систем
Способность применять в исследовательской и прикладной деятельности современные языки программирования и языки манипулирования данными, операционные системы, пакеты программ и т.д.	ИК-М7.5 пми	Студент уверенно составляет программы на языке высокого уровня для проведения вычислений по применяемым моделям компьютерной лингвистики	Домашние задания, ориентированные на программную реализацию вычислительных моделей обработки текста, и проведение с их помощью экспериментов
Способность публично представлять результаты профессиональной деятельности (в том числе с использованием информационных технологий)	ИК-М2.5 пми	Студент способен провести анализ и представить в виде сжатого доклада материал научных статей по автоматической обработке текстов	Подготовка и проведение (в форме презентации) научного доклада по одной из актуальных проблем компьютерной лингвистики

5. Место дисциплины в структуре образовательной программы

Настоящая учебная дисциплина является дисциплиной по выбору в магистерской программе «Науки о данных», специализация «Интеллектуальные системы и структурный анализ», для подготовки магистра по направлению 01.04.02 «Прикладная математика и информатика».

Для освоения дисциплины предполагаются базовые знания по таким разделам математики и информатики, как «Дискретная математика», «Теория вероятностей и математическая статистика», «Информатика и программирование», «Алгоритмы и структуры данных» – соответствующие дисциплины входят в программу обучения бакалавра по направлению 010400.62 «Прикладная математика и информатика».

Знания по дисциплине «Компьютерная лингвистика и анализ текста» предполагаются в дальнейшем при изучении нескольких дисциплин указанной магистерской программы, включая дисциплины:

- Методы машинного обучения и разработки данных,
- Современные методы анализа данных.

6. Тематический план дисциплины «Компьютерная лингвистика и анализ текста»

№	Название темы	Всего часов по дисциплине	Аудиторные часы		Самостоятельная работа
			Лекции	Практ.зан.	
1	Введение	9	3	0	6
2	Морфологические модели	24	4	4	16
3	Методы синтаксического анализа	33	5	6	22
4	Статистические модели	26	4	6	16
5	Модели дискурса и семантики	24	4	4	16
6	Построение и применение лингвистических ресурсов	32	4	4	24
7	Разработка приложений КЛ	42	6	6	30
	Итого	190	30	30	130

7. Формы контроля знаний студентов

Курс «Компьютерная лингвистика и анализ текста» читается в 3 и 4 модуле.

Тип контроля	Форма контроля	Параметры
Текущий контроль	Контрольная работа	Письменная работа 60 минут
	Домашнее задание	Выдается для выполнения в течение 2 недель
Итоговый контроль в 4 модуле	Экзамен	Письменная работа 80 минут

Критерии оценки знаний

На текущем и итоговом контроле студент должен продемонстрировать владение основными понятиями по пройденным темам дисциплины.

Текущий контроль включает письменную контрольную, состоящую из нескольких вопросов и задач по пройденному материалу, а также домашние задания на применение моделей компьютерной лингвистики и лингвистических ресурсов (в некоторых заданиях требуется программная реализация моделей и проведение экспериментов). Одно из домашних заданий – подготовка доклада (презентации) по определенной актуальной проблеме из области автоматической обработки текста на естественном языке.

Итоговый контроль проводится в форме письменного экзамена, включающего несколько вопросов и задач по темам дисциплины.

Порядок формирования оценок по дисциплине

Накопленная оценка (за текущий контроль) $O_{накопленная}$ учитывает оценку $O_{к/р}$ за письменную контрольную работу и оценку $O_{д/з}$ самостоятельной работы студентов при

выполнении домашних заданий по текущим темам дисциплины, которая рассчитывается по десятибалльной шкале как средняя оценка всех домашних заданий и округляется до целого числа арифметическим способом. ***Накопленная оценка*** рассчитывается согласно следующей формуле: $O_{накопленная} = 0{,}5\ O_{к/р} + 0{,}5 \cdot O_{д/з}$

Результирующая оценка по данной учебной дисциплине выставляется по формуле:

$$O_{дисциплина} = 0{,}7 \cdot O_{накопленная} + 0{,}3 \cdot O_{экзамен}$$

с окрулением до целого числа арифметическим способом, где $O_{экзамен}$ – оценка экзамена в конце 4 модуле.

8. Содержание программы по темам

Тема 1. Введение

1. Компьютерная лингвистика и автоматическая обработка текстов на естественном языке (ЕЯ): основные задачи и история развития. Междисциплинарный характер направления, связь со смежными научными дисциплинами.

2. Особенности ЕЯ, понятия языкового знака и языковой системы. Принципиальные отличия естественных и искусственных (формальных) языков: открытость, избыточность, нестандартная сочетаемость, асимметрия знаков и смыслов.

3. Уровни языковой системы (от фонетики до дискурса), их взаимосвязь. Основные единицы текста. Уровень фонем и символов. Синтаксический и морфологический уровни. Лексическая система. Словоформы и лексемы.

4. Понятие модели в компьютерной лингвистике. Основные уровни обработки текста и виды моделей. Модель «Смысл-Текст». Лингвистический процессор и лингвистические ресурсы (компьютерные словари и тезаурусы, грамматики, корпуса текстов).

Основная литература

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И. и др. – М.: МИЭМ, 2011.
2. Леонтьева Н. Н. Автоматическое понимание текстов: Системы, модели, ресурсы: Учебное пособие – М.: Академия, 2006.

Дополнительная литература

1. Белоногов Г.Г. Компьютерная лингвистика и перспективные информационные технологии. – М.: Русский мир, 2004.
2. Касевич В.Б. Элементы общей лингвистики. — М., Наука, 1977.
3. Bolshakov, I.A., Gelbukh A. Computational Linguistics. Models, Resources, Applications. Mexico, IPN, 2004.
4. The Oxford Handbook on Computational Linguistics. R. Mitkov (Ed.). Oxford University Press, 2005.

Тема 2. Морфологические модели

1. Основные понятия морфологических моделей: морфема, аффикс, корень, основа, флексия. Словоформа и лексема. Основа и псевдооснова. Словоизменительная парадигма, флективный класс. Морфологические модели на базе словаря. Словари основ и словари словоформ. Особенности русской морфологии.

2. Морфемный состав слова. Виды морфем, понятие алломорфа. Морфотактики.

3. Морфологический анализ и синтез. Основные подходы морфологического анализа: анализ на базе словаря, бессловарный анализ. Виды морфоанализа: лемматизация, стемминг, полный морфоанализ. Программные модули автоматического морфологического анализа.

4. Графематический анализ и сегментация текста. Виды сегментации. Токенизация. Проблемы графематического анализа, технологии его реализации на базе конечных автоматов и регулярных выражений.

Основная литература

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И. и др. – М.: МИЭМ, 2011.
2. Васильев В. Г., Кривенко М. П. Методы автоматизированной обработки текстов. – М.: ИПИ РАН, 2008.
3. Леонтьева Н. Н. Автоматическое понимание текстов: Системы, модели, ресурсы: Учебное пособие – М.: Академия, 2006.

Дополнительная литература

1. Болховитянов А.В., Гусев А.В., Чеповский А.М. Морфологические модели компьютерной лингвистики: учеб. пособие – М. МГУП, 2010.
2. Лингвистический энциклопедический словарь / Гл. ред. В.Н.Ярцева, 2-ое изд., дополненное – М.: Научное издательство "Большая Российская энциклопедия", 2002.
3. The Oxford Handbook on Computational Linguistics. R. Mitkov (Ed.). Oxford University Press, 2005.

Тема 3. Методы синтаксического анализа

1. Различные подходы к анализу синтаксиса предложений ЕЯ. Основная задача синтаксического анализа. Синтаксические деревья: деревья непосредственных составляющих и деревья зависимостей. Синтаксические связи. Проективность предложений. Понятия синтаксического предиката, валентности и актанта, модели управления. Теория синтаксических групп Гладкого.

2. Методы синтаксического разбора на базе контекстно-свободных (КС) грамматик. Нисходящий и восходящий разбор. Применение автоматов и преобразователей с конечным числом состояний (Finite State Transducers). Синтаксический анализ на основе грамматик зависимостей. Синтаксические парсеры для английского и русского языков.

3. Частичный синтаксический анализ. Словосочетания и их основные синтаксические типы. Задача синтаксической сегментации текста. Выделение синтаксических групп.

Основная литература

1. Васильев В. Г., Кривенко М. П. Методы автоматизированной обработки текстов. – М.: ИПИ РАН, 2008.
2. Леонтьева Н. Н. Автоматическое понимание текстов: Системы, модели, ресурсы: Учебное пособие – М.: Академия, 2006.
3. The Oxford Handbook on Computational Linguistics. R. Mitkov (Ed.). Oxford University Press, 2005.

Дополнительная литература

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И. и др. – М.: МИЭМ, 2011.
2. Апресян Ю.Д. и др. Лингвистическое обеспечение системы ЭТАП-2. М.: Наука, 1989.
3. Гладкий А.В. Синтаксические структуры естественного языка в автоматизированных системах общения. — М., Наука, 1985.
4. Касевич В.Б. Элементы общей лингвистики. — М., Наука, 1977.
5. Bolshakov, I.A., Gelbukh A. Computational Linguistics. Models, Resources, Applications. Mexico, IPN, 2004.
6. Jurafsky D., Martin J. Speech and Language Processing. An Introduction to Natural Language Processing, Comp. Linguistics and Speech Recognition. Prentice Hall, 2000.

Тема 4. Статистические модели

1. Статистические характеристики текстов ЕЯ. Статистика встречаемости букв и буквосочетаний: биграмм, триграмм, N-грамм. Статистика N-грамм для слов текста. Статистические языковые модели. Цепи Маркова и их применение. Приложения статистических моделей.

2. Статистика встречаемости сочетаний слов. Типы словосочетаний по фразеологичности. Лексические функции модели «Смысл-Текст». Понятие коллокации. Устойчивые словосочетания, методы их автоматического извлечения из текстов на базе статистики. Меры ассоциации и устойчивости.

Основная литература

1. Manning, Ch. D., H. Schütze. Foundations of Statistical Natural Language Processing. MIT Press, 1999.
2. The Oxford Handbook on Computational Linguistics. R. Mitkov (Ed.). Oxford University Press, 2005.

Дополнительная литература

1. Пиотровский Р.Г. , Бектаев К.Б., Пиотровская А.А. Математическая лингвистика. – М.: Высшая школа, 1977.
2. Чатуев М.Б., Чеповский А.М. Частотные методы в компьютерной лингвистике: учебное пособие – М. МГУП, 2011.
3. Jurafsky D., Martin J. Speech and Language Processing. An Introduction to Natural Language Processing, Comput. Linguistics and Speech Recognition. Prentice Hall, 2000.

Тема 5. Модели дискурса и семантики

1. Характеристики связного текста (дискурса): тематическая связность, риторическая связность, лексическая связность, референциальная связность. Целостность и связность. Анафорические ссылки, кореференция, лексические повторы, дискурсивные слова. Тематическая и композиционная структура текста. Сверхфразовые единства.

2. Моделирование свойств связного текста. Построение лексических цепочек, автоматическое разрешение референции. Композиционные и дискурсивные особенности текстов разных жанров и стилей, их учет при обработке текстов.

3. Модели представления семантики. Семантико-синтаксическая модель управления слов и семантические роли. Типизированные структуры. Лингвистический ресурс FrameNet: состав, принципы построения. Задача разметки семантических ролей.

Основная литература

1. Леонтьева Н. Н. Автоматическое понимание текстов: Системы, модели, ресурсы: Учебное пособие – М.: Академия, 2006.
2. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – М.: Изд-во Московского университета, 2011.

Дополнительная литература

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И. и др. – М.: МИЭМ, 2011.
2. Ван Дейк Т.А., Кинч В. Стратегия понимания связного текста.// Новое в зарубежной лингвистике. Вып. XXIII — М., Прогресс, 1988, с. 153-211.
3. Зубов А.В., Зубова И.И. Основы искусственного интеллекта для лингвистов. – М., Логос, 2007.
4. Кобозева И.М. Лингвистическая семантика. – М., 2009.
5. Кронгауз М.А. Семантика. – М.: Издательский центр «Академия», 2005.

6. Лингвистический энциклопедический словарь / Гл. ред. В.Н.Ярцева, 2-ое изд., дополненное – М.: Научное издательство "Большая Российская энциклопедия", 2002.
7. Плунгян В.А. Введение в грамматическую семантику: грамматические значения и грамматические системы языков мира. – М.: РГГУ, 2011.

Тема 6. Построение и применение лингвистических ресурсов

1. Система понятий и терминов как основа описания предметной области. Синонимия и лексическая многозначность. Смысловые (парадигматические) отношения лексических единиц. Лексический ресурс WordNet: состав, принципы построения. Лингвистические онтологии, ресурс EvroNet.

2. Методы извлечения терминологических слов и словосочетаний из текстов. Способы оценки качества извлечения. Автоматизация выявления терминологических связей: извлечение синонимов терминов, установление родовидовых отношений. Автоматизация построения таксономий.

4. Коллекции и корпуса текстов. Корпусная лингвистика. Типы и характеристики корпусов, виды разметки текстов. Параллельные и псевдопараллельные корпуса текстов, их применение.

Основная литература

1. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – М.: Изд-во Московского университета, 2011.
2. Biber, D., Conrad S., and Reppen D. Corpus Linguistics. Investigating Language Structure and Use. Cambridge University Press, Cambridge, 1998.
3. Hirst, G. Ontology and the Lexicon. In.: Handbook on Ontologies in Information Systems. Berlin, Springer, 2003.
4. Word Net: an Electronic Lexical Database. /Edit. by Christiane Fellbaum. Cambridge, MIT Press, 1998.

Дополнительная литература

1. Лингвистический энциклопедический словарь / Гл. ред. В.Н.Ярцева, 2-ое изд., дополненное – М.: Научное издательство "Большая Российская энциклопедия", 2002.
2. Кобозева И.М. Лингвистическая семантика. – М., 2009.
3. Национальный Корпус Русского Языка. http://ruscorpora.ru
4. Manning, Ch. D., H. Schütze. Foundations of Statistical Natural Language Processing. MIT Press, 1999.
5. The Oxford Handbook on Computational Linguistics. R. Mitkov (Ed.). Oxford University Press, 2005.

Тема 7. Разработка приложений

1. Обзор приложений компьютерной лингвистики (КЛ). Подходы к разработке приложений: подход, основанный на знаниях (правилах), и подход, основанный на машинном обучении. Сравнение подходов. Основные показатели качества решения: точность, полнота, F-мера.

2. Машинный перевод (МП). Лингвистические стратегии машинного перевода и поколения систем МП. Автоматический перевод, основанный на правилах. Интерлингва. Статистический машинный перевод. Оценки качества машинного перевода.

3. Генерация текстов документов. Стратегии синтеза текста. Генерация многоязыковых руководств пользователя (инструкций) по формальному описанию проблемной области.

4. Извлечение информации из текстов (Information Extraction): объектов (именованных сущностей), их связей, фактов. Лингвистические шаблоны и их использование. Оценки

качества извлечения информации. Инструментальные программные средства: система GATE как типичная среда построения приложений для извлечения информации.

5. Автоматический анализ тональности текстов и извлечение мнений из текстов. Автоматизированное построение базы знаний для систем анализа тональности.

6. Реферирование и аннотирование документов. Типы аннотаций: индикативная и информативная аннотация, аннотация по запросу, аннотация нескольких документов.

7. Приложения КЛ, основанные на статистике и векторной модели текста. Классификация и кластеризация документов, рубрикация. Распознавание авторства текстов, дубликатов документов. Выбор признаков и метрик.

Основная литература

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И. и др. – М.: МИЭМ, 2011.
2. Леонтьева Н. Н. Автоматическое понимание текстов: Системы, модели, ресурсы: Учебное пособие – М.: Академия, 2006.
3. Pang Bo, Lee L. Opinion Mining and Sentiment Analysis. In: Foundations and Trends® in Information Retrieval. Now Publishers, 2008.

Дополнительная литература

1. Барсегян А.А. и др. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP – 2-е изд. – СПб.: БХВ-Петербург, 2008.
2. Васильев В. Г., Кривенко М. П. Методы автоматизированной обработки текстов. – М.: ИПИ РАН, 2008.
3. Маккьюин К. Дискурсивные стратегии для синтеза текста на естественном языке // Новое в зарубежной лингвистике. Вып. XXIV. М.: Прогресс, 1989, с.311-356.
4. Brown P., Pietra S., Mercer R., Pietra V. The Mathematics of Statistical Machine Translation. // Computational Linguistics, Vol. 19(2): 263-311. 1993.
5. Bolshakov, I.A., Gelbukh A. Computational Linguistics. Models, Resources, Applications. Mexico, IPN, 2004.
6. The Oxford Handbook on Computational Linguistics. R. Mitkov (Ed.). Oxford University Press, 2005

9. Образовательные технологии

В преподавании данной дисциплины сочетаются:

- лекции в форме презентаций (которые затем высылаются студентам для их самостоятельной работы);
- практические занятия для обсуждения особенностей моделей компьютерной лингвистики и решения задач по отдельным темам;
- домашние задания на применение изученных моделей компьютерной лингвистики.

10. Оценочные средства для текущего и итогового контроля

Примеры домашних заданий

1) Для заданного слова русского языка найти и сравнить его толкования в различных толковых словарях, а также в Национальном корпусе русского языка.
2) Для заданного набора словосочетаний отобрать явные термины и явные нетермины, объяснить принятые решения.

Вопросы для оценки качества освоения дисциплины

Тема 1.

1. С какими научными дисциплинами связана область компьютерной лингвистики?
2. Перечислите основные отличия естественных языков от искусственных.
3. В чем суть явления полисемии? омонимии? Приведите примеры.
4. Перечислите основные уровни (подсистемы) языковой системы.
5. В чем особенности компьютерных моделей естественного языка?

Тема 2.

6. Охарактеризуйте понятие лексемы.
7. Что такое морфема? аффикс? Какие виды аффиксов вы знаете?
8. Чем основа слова отличается от корня? Приведите примеры.
9. Что такое словоизменительная парадигма?
10. В чем заключается лемматизация?
11. Назовите основные стратегии морфологического анализа.
12. Приведите пример морфологической омонимии.

Тема 3.

13. Что такое синтаксическое дерево?
14. В чем особенность деревьев составляющих? Приведите пример.
15. В чем особенность деревьев зависимостей? Приведите пример.
16. Чем проективное предложение ЕЯ отличается от непроективного?
17. Что такое валентность? Актант? Приведите примеры.
18. Какие методы и алгоритмы анализа контекстно-свободных языков вы знаете?
19. В чем состоит синтаксическая сегментация текста?
20. Какие синтаксические типы словосочетаний вы знаете?

Тема 4.

21. Какие статистические характеристики рассчитываются в статистических моделях?
22. Что такое N-грамма?
23. Объясните понятие устойчивого словосочетания.
24. Что такое мера взаимной информации Mutual Information?
25. Какие статистические меры применяются для извлечения коллокаций?

Тема 5.

26. Назовите отличительные характеристики связного текста.
27. Что такое анафорическая ссылка?
28. Поясните понятие сверхфразового единства.
29. Объясните понятие лексической цепочки. Приведите примеры.
30. Что такое тематическая структура текстов? Риторическая структура?
31. Укажите принципы автоматического разрешения референции.
32. Какие модели семантики текста вы знаете?
33. В чем состоит задача разметки семантических ролей?

Тема 6.

34. Что такое термин? Приведите примеры.
35. Назовите основные свойства терминов.
36. Какие свойства родовидовых (таксономических отношений) вы знаете?
37. Укажите принципы установления родовидовых (таксономических) отношений.
38. Какие подвиды отношения часть-целое вы можете назвать?
39. Какие отличительные признаки корпуса текстов вы можете назвать?
40. Что такое параллельный и псевдопараллельный корпус?

Тема 7.

41. Назовите и охарактеризуйте типичные приложения компьютерной лингвистики.
42. В чем особенности задачи извлечения информации из текстов?
43. Укажите основные стратегии машинного перевода. Что такое интерлингва?

44. В каких прикладных задачах применяется генерация текста?
45. Охарактеризуйте понятие лингвистического шаблона.

Примеры вопросов и задач на экзаменах

Воп росы:

1. Что является результатом полного морфологического анализа словоформы? Поясните на примере конкретной словоформы.
2. В чем отличие синтаксических деревьев непосредственно составляющих от синтаксических деревьев зависимостей?
3. Что такое непроективность предложения? Слабая непроективность? Привести собственные примеры.
4. Что такое валентность? Актант? Слово-предикат? Проиллюстрировать на примере.
5. Какие бывают виды сегментации текста? Охарактеризуйте синтаксическую сегментацию.
6. Укажите типы лексических отношений.
7. Что такое семантические классы слов?
8. Охарактеризуйте особенности описания семантических ролей в проекте FrameNet.
9. Какие психолингвистические предположения легли в основу создания WordNet?
10. Что такое лингвистические онтологии? В чем их особенность?
11. Назовите классы словосочетаний по степени фразеологичности, приведите примеры.
12. Укажите лингвистические критерии извлечения коллокаций.
13. Что такое мера ассоциации? Для чего применяются меры? Приведите примеры мер.
14. Укажите формулу подсчета и особенности применения меры Mutual Information для извлечения коллокаций.
15. Статистический машинный перевод: особенности технологии.
16. Особенности задачи автоматической генерации текстов, этапы генерации.

Задачи :

1) Для заданной фразы построить возможные синтаксические деревья зависимостей.
2) Составить семантико-синтаксическую модель управления заданного слова.

11. Учебно-методическое и информационное обеспечение дисциплины

Базовый учебник – ридер «Автоматическая обработка текста», составленный по следующим источникам:

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И. и др. – М.: МИЭМ, 2011.
2. Васильев В. Г., Кривенко М. П. Методы автоматизированной обработки текстов. – М.: ИПИ РАН, 2008.
3. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – М.: Изд-во Московского университета, 2011.
4. The Oxford Handbook on Computational Linguistics. R. Mitkov (Ed.). Oxford University Press, 2005.

Дополнительная литература

1. Белоногов Г.Г. Компьютерная лингвистика и перспективные информационные технологии. – М.: Русский мир, 2004.
2. Болховитянов А.В., Гусев А.В., Чеповский А.М. Морфологические модели компьютерной лингвистики: учеб. пособие – М. МГУП, 2010.

3. Ван Дейк Т.А., Кинч В. Стратегия понимания связного текста.// Новое в зарубежной лингвистике. Вып. XXIII — М., Прогресс, 1988, с. 153-211.
4. Гладкий А.В. Синтаксические структуры естественного языка в автоматизированных системах общения. — М., Наука, 1985.
5. Касевич В.Б. Элементы общей лингвистики. — М., Наука, 1977.
6. Кобозева И.М. Лингвистическая семантика. – М., 2009.
7. Кронгауз М.А. Семантика. - М.: Издательский центр «Академия», 2005.
8. Маккьюин К. Дискурсивные стратегии для синтеза текста на естественном языке // Новое в зарубежной лингвистике. Вып. XXIV. М.: Прогресс, 1989, с.311-356.
9. Леонтьева Н. Н. Автоматическое понимание текстов: Системы, модели, ресурсы: Учебное пособие – М.: Академия, 2006.
10. Лингвистический энциклопедический словарь / Гл. ред. В.Н.Ярцева, 2-ое изд., дополненное – М.: Научное издательство "Большая Российская энциклопедия", 2002.
11. Пиотровский Р.Г. , Бектаев К.Б., Пиотровская А.А. Математическая лингвистика. – М.: Высшая школа, 1977.
12. Чатуев М.Б., Чеповский А.М. Частотные методы в компьютерной лингвистике: учебное пособие – М. МГУП, 2011.
13. Biber, D., Conrad S., and Reppen D. Corpus Linguistics. Investigating Language Structure and Use. Cambridge University Press, Cambridge, 1998.
14. Bolshakov, I.A., Gelbukh A. Computational Linguistics. Models, Resources, Applications. Mexico, IPN, 2004.
15. Jurafsky D., Martin J. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Prentice Hall, 2000.
16. Manning, Ch. D., H. Schütze. Foundations of Statistical Natural Language Processing. MIT Press, 1999.
17. Word Net: an Electronic Lexical Database. /Edit. by Christiane Fellbaum. Cambridge, MIT Press, 1998.

13. Материально-техническое обеспечение дисциплины

Для лекций и семинарских занятий по темам дисциплины используется проектор и компьютеры с выходом в сеть Интернет.

Автор программы:_____/ Большакова Е.И. /