Программа учебной дисциплины «Компьютерная лингвистика»

Утверждена
Академическим советом ООП
Протокол № 11 от «28» июня 2018 г.

Автор	Толдова С.Ю.
Число кредитов	5
Контактная работа (час.)	64
Самостоятельная работа (час.)	126
Курс	1
Формат изучения дисциплины	без использования онлайн курса

I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Целями освоения дисциплины «Компьютерная лингвистика» являются овладение студентами основными методами автоматической обработки текста на разных уровнях лингвистического анализа.

В результате освоения дисциплины студент должен:

знать:

- современные задачи компьютерной лингвистики;
- основные формальные модели, лежащие в основе различных модулей автоматической обработки текста;
- необходимые этапы морфологического анализа и проблемы, возникающие при моделировании каждого из этапов;
- современные алгоритмы, используемые для построения автоматического синтаксического анализа;
- наиболее известные доступные для свободного использования компоненты автоматического анализа, в том числе синтаксические и морфологические парсеры;
- принципы оценки качества таких систем;

уметь:

- создавать модули первичной обработки текста;
- строить формальную модель морфологии для создания системы автоматического морфологического анализа;
- проводить оценку качества систем автоматического морфологического, синтаксического и семантического анализа;
- использовать соответствующие модули в различных приложениях;

владеть:

- разработки программ первичной обработки текста;
- использования систем автоматического морфологического анализа;
- тестирования систем морфологического и синтаксического анализа.

Изучение дисциплины «Компьютерная лингвистика» базируется на следующих дисциплинах:

- Введение в лингвистику (первый и второй курс) программы подготовки бакалавра
- Теория языка
- Линейная алгебра и математический анализ программы подготовки бакалавра

- Программирование и компьютерные инструменты лингвистического исследования программы подготовки бакалавра
- Дискретная математика программы подготовки бакалавра
- Программирование и компьютерные инструменты лингвистического исследования программы подготовки бакалавра
- Иностранный язык

Для освоения учебной дисциплины студенты должны владеть следующими знаниями и компетенциями:

- владеть базовыми представлениями о грамматических категориях и анализе языковых единиц;
- владеть базовыми знаниями в области теории алгоритмов и основ математики
- владеть базовыми знаниями в области теории вероятностей и статистики;
- уметь читать научные работы и технические описания на английском языке.

Основные положения дисциплины должны быть использованы в дальнейшем при изучении следующих дисциплин:

- компьютерной лингвистики программы подготовки магистров второго курса
- научно-исследовательский семинар
- машинное обучение
- производственная практика по специальности
- подготовка и защита выпускной квалификационной работы

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Тема 1. Введение в компьютерную лингвистику

Введение в компьютерную лингвистику. Задачи компьютерной лингвистики. Модель информационного поиска. Новостная агрегация и рубрикация. Извлечение информации из текста. Основные типы ресурсов. Основные формальные модели: конечные автоматы, контекстно-свободные грамматики

Свойства естественного языка, создающие сложности для автоматической обработки: омонимия, отсутствие взаимооднозначного соответствия между формой и смыслом. Цепочка обработки: основные этапы обработки. Основные платформы и пакеты для разработки систем АОТ.

Тема 2. Первичная обработка текста.

Графематический анализ. Сегментация текста. Проблемы токенизации: токены; Стоп-слова; обработка специальных символов; обработка слов с дефисом. Типизация токенов. Оффсеты. Сегментация на предложения. Сегментация текста в библиотеке NLTK.

Сегментация на слова в беспробельных языках: алгоритмs MinMatch, MaxMatch

Тема 3. Ресурсы

Регулярные выражения. Гигакорпуса Arranea Поиск в корпусе. Квантитативные характеристики языковых единиц. Разметка корпуса.

Тема 4. Модель информационного поиска. Векторизация текста

Модель информационного поиска. Модель мешка слов. Индексация. Матрица терм-документ. N-грамы. tf.idf. Оценка качества. Векторная модель документа, векторная модель слова. Поиск похожих текстов. Косинусная мера близости.

Векторизация текстов в библиотеке scikit-learn

Тема 5. Распознавание языка

Частотный анализ лексики и ключевые слова. N-грамы. Методы автоматического определения языков

Тема 6. Методы исправления ошибок

Расстояние Левенштейна. Алгоритм Питера Норвига.

N-грамные модели при исправлении ошибок

Тема 4. Автоматический морфологический анализ.

Введение в автоматический морфологический анализ. Постановка задачи. Основные типы морфологической обработки.

Явления неконкатенативной морфологии. Конечные автоматы и конченые преобразователи. Моделирование морфонологических процессов: конечные преобразователи / суффиксные деревья. Сегментная и автосегментная морфология. Среда для проектирования системы морфологического анализа на примере тюркских и финно-угорских языков с большим количеством морфонологических чередований. Модель фрагмента морфологии языка X с использованием одной из платформ лдя реализации конечных преобразователей

Проблемы морфологической неоднозначности. Методы дизамбигуации. Языковые модели. Скрытые марковские модели. Алгоритм Витерби. Метода максимальной энтропии. Метод случайных полей.

Оценка качества частеречного тегера: практикум.

Тема 5. Автоматический синтаксический анализ

Основные модели автоматического синтаксического анализа: непосредственные составляющие, зависимости. Контекстно-свободные грамматики. Унификационные грамматики.

Синтаксический анализ: основные проблемы автоматического анализа (омонимия, типичные случаи синтаксической омонимии, синтаксические нули).

Зависимостные грамматики. Основные проблемы анализа в терминах зависимостей. Разметка деревьев. Алгоритмы анализа в терминах зависимостей.

Универсальные зависимости (UD): основные стандарты морфологической и синтаксической разметки в терминах UD. Запуск системы синтаксического анализа в терминах UD (UD-pipe). Обучение системы синтаксического анализа в терминах зависимостей.

Использование векторных моделей и нейронных сетей в автоматическом синтаксическом анализе.

III. ОЦЕНИВАНИЕ

Оценки по всем формам текущего контроля выставляются по 10-ти балльной шкале.

Накопленная оценка складывается следующим образом.

- Текущие тесты - 15%

В течение семестра проводятся короткие тесты, оценивающие теоретические знания по предыдущей теме (вопросы с выбором варианта ответа или с кратким ответом).

- Домашние задания - 40%

В течение семестра предлагаются практические задания по реализации алгоритмов и методов, обсуждаемых на занятиях

- Проектная работа – 45%

В рамках курса готовится проект по тестированию одной из автоматичсеких систем морфологического или синтаксического анализа для русского языка.

- Итоговый экзамен – 40%

В качестве итогового контроля освоения дисциплины предлагается письменный тест.

Накопленная оценка по 10-ти балльной шкале вычисляется по формуле:

$$O_{\text{накопленная}} = 0{,}15 \cdot O_{\text{текущие тесты}} + 0{,}4 \cdot O_{\text{домашнее задание}} + 0{,}45 \cdot O_{\text{контрольная}}$$

Оценка за итоговый контроль в форме экзамена $O_{\text{экзамен}}$ выставляется по 10-балльной шкале. Проект включает подготовку корпуса, разработку программы предварительного анализа текста, тестирование морфологического и синтаксического анализатора.

Тестирование программы предварительного анализа текста и морфологических анализаторов проходит в формате Форума по оценке систем автоматической обработке текста. Командам выдается тестовый и эталонный корпус. Каждая команда проводит оценку точности и полноты, а также функциональное тестирование соответствующей программы.

Результирующая оценка вычисляется по следующей формуле:

$$O_{\text{итоговый}} = 0{,}4 \cdot O_{\text{экзамен}} + 0{,}6 \cdot O_{\text{накопленная}},$$

Способ округления накопленной оценки текущего контроля: арифметический.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

Оценочные средства для текущего контроля студента

Примеры домашних заданий

- Создайте список наиболее частотных терминов вашего корпуса.
- Разбейте текст на токены и предложения. Составьте список необходимых для сегментации классов символов, типов токенов; составьте список сложных случаев, предложите решение.
- Протестируйте систему сегментации текста
- Разработайте программу для распознавания n-языков
- Постройте конечный автомат и конечный преобразователь для описания фрагмента морфологического анализа для предложенного языка
- Запустите систему морфологического анализа. Проведите морфологическую дизамбигуацию некоторого текста. Какие типы омонимии вам встретились. Оцените качество предсказаний системы.
- Постройте контекстно-свободную грамматику для анализа некоторой синтаксической конструкции
- Предложите синтаксическую разметку предложения, основанную на деревьях зависимости
- Тестирование системы синтаксического анализа в терминах универсальных зависимостей

Мини-тесты по материалам лекций и прочитанной литературе

Пример вопросов мини-теста:

1. Назовите критерии определения вершин в грамматике зависимостей, приведите примеры. В каких случаях в разметке UD, принципы выделения вершины не соответствуют теоретическим принципам. Какие аргументы.
2. Приведите два разных представления данных для морфологического анализа на примере анализа словоформы "городка"
3. На каком допущении относительно вероятности тега в цепочке тегов базируется метод дизамбигуации, основанный на скрытых марковских моделях.
4. Дана словоформа "данные". Определите, в чем проблема ее лемматизации

Оценочные средства для промежуточной аттестации

1. Какие типы лингвистических данных вам известны?
2. Принципы работы морфологических парсеров
3. Конечные автоматы и конечные преобразователи в морфологическом анализе
4. Методы снятия морфологической неоднозначности
5. Основания оценки качества автоматического морфологического разбора
6. Технология shallow parcing
7. Технология chunking
8. Принципы работы синтаксических парсеров
9. Основания оценки качества автоматического синтаксического разбора

Пример теста:

Пример контрольной работы

Часть 1. Без использования материалов

1. Перечислите основные направления в области
    a. извлечения информации / знаний из текста.

   Кратко охарактеризуйте основную задачу / основные задачи направления. Назовите 2-3 термина, связанные с этими направлениями с кратким пояснением
2. За что отвечает этап идиоматизации. Приведите примеры различных типов случаев, которые этот этап «обслуживает»
3. Что является наблюдаемой переменной в скрытой марковской модели, применяемой для частеречной аннотации

   На каких допущениях относительно лексической вероятности (грамматических тегов для конкретной словоформы) базируется метод дизамбигуации, основанный на скрытых марковских моделях. Приведите пример ситуаций, когда это основное не работает.
4. Что является синтаксическими единицами, синтаксическими отношениями в представлении синтаксической структуры в терминах непосредственных составляющих
5. Приведите пример правил для синтаксического анализа в терминах КС- грамматики, которые позволяют порождать предложения

   Я вижу лес

   Ты видишь лес и т.п.

   Но запрещают предложения типа
   - Я видишь лес
   - Ты видим лес и т.п.

Часть 2.

6. Найдите в НКРЯ три ошибки при делении на предложения. Ответ прокомментируйте. (Нужно придумать такие запросы, которые бы вам искали ошибки определенного типа)
7. Вычислите tf.idf для слов *медвежонок*, *что, ворона* в тексте
   Сергей Козлов. Как Ёжик с Медвежонком спасли Волка // «Мурзилка», 2003
   для коллекции НКРЯ «Детская литература»
8. Методы дизамбигуации:
   Приведите пример 2-х правил (patch-a) в методе Эрика Брилла, которые можно вывести из следующего фрагмента
   Золотой стандарт:
   The fly can fly
   Det N V V
   Первичная аннотация
   Det Verb Verb N
   Приведите пример миникорпуса (4 предожения), на котором одно из полученных правил увеличит количество ошибок, а не уменьшит?
9. Приведите глубинное (lexical) – промежуточное (intermediate) – поверхностное представление (Surface) для словоформы
   дымка (в двухуровневой морфологии)
10. Дан набор словоформ. Предложить 3 варианта представления морфологических данных для данной лексемы и данной подпарадигмы: глагол *сидеть* в настоящем времени
11. Вычислите вероятность прилагательного в контексте предшествующего ему указательного местоимения и сравните ее с вероятностью глагола в контексте предшествующего ему указательного местоимения
12. Дан Трибанк из 4-х предложений, постройте по ним КС-грамматику и переведите ее в нормальную форму Хомского
    Дано предложение и грамматика. Представьте шаги анализа данного предложения при применении алгоритма Кока-Янгера-Касами
    1) Вася читает мою книгу
    2) Напиши какое-нибудь письмо
    3) Этот веселый мальчик идет
    4) Он любит читать всякие книги

Предложение для разбора: *Они используют эти виды стали.*

V. РЕСУРСЫ

5.1 Основная литература

Jurafsky, D., Martin J. H. Speech and Language Processing, 3 издание. 2018
https://web.stanford.edu/~jurafsky/slp3/

Автоматическая обработка текстов на естественном языке и компьютерная лингвистика / Е.И. Большакова, Э.С.Клышинский, Д.В. Ландэ, А.А.Носков, О.В. Пескова, Е.В. – Ягунова М.: МИЭМ, 2011 г. – URL: http://window.edu.ru/catalog/pdf2txt/465/78465/59324.

5.2 Дополнительная литература

Perkins J. Python Text Processing with NLTK 2.0 Cookbook: Over 80 Practical Recipes for Using Python's NLTK Suite of Libraries to Maximize Your Natural Language Processing Capabilities. / Jacob Perkins ed. – Packt Publishing. – 2010. – URL: https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=1126730. – ЭБС ProQuest Ebook Central - Academic Complete.

5.3 Программное обеспечение

№ п/п	Наименование	Условия доступа
1.	Microsoft Windows 10 Microsoft Windows 8.1 Professional RUS	*Из внутренней сети университета (договор)*
2.	Microsoft Office Professional Plus 2010	*Из внутренней сети университета (договор)*
3.	Python 3	*Свободный*
4.	Ubuntu 18	*Свободный*
5.	NLTK	*Свободный*
6.	Библиотеки numpy, scipy, pandas, matplotlib, plotly, scikit-learn, gensim, xgboost	*Свободный*

5.4

5.5 Профессиональные базы данных, информационные справочные системы, интернет-ресурсы (электронные образовательные ресурсы)

№ п/п	Наименование	Условия доступа / скачивания
	Профессиональные базы данных, информационно-справочные системы	
1	ЭБС ProQuest Ebook Central - Academic Complete	URL: https://www.proquest.com/libraries/academic/
	Интернет-ресурсы	
1	Единое окно к образовательным ресурсам [Электронный ресурс]	URL: http://window.edu.ru
2	http://informationretrieval.org	свободный доступ
3	https://web.stanford.edu/~jurafsky/slp3/	свободный доступ

5.6 Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают исполь-

зование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

- ПЭВМ с доступом в Интернет (операционная система, офисные программы, антивирусные программы);
- мультимедийный проектор с дистанционным управлением.

Учебные аудитории для семинарских и самостоятельных занятий по дисциплине не требуют специального технического оснащения