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

Факультет гуманитарных наук
Школа лингвистики

Рабочая программа дисциплины
Автоматическая обработка естественного языка

для образовательной программы "Фундаментальная и компьютерная лингвистика"
направления подготовки 45.03.03. Фундаментальная и прикладная лингвистика
уровень бакалавр

Разработчик(и) программы
Толдова С.Ю, к.ф.н., stoldova@hse.ru

Одобрена на заседании Школы лингвистики ФГН «30» мая 2016 г.
Руководитель Школы лингвистики Е.В. Рахилина

Рекомендована Академическим советом образовательной программы
«01» июня 2016 г., № протокола 10

Утверждена «01»июня 2016 г.
Академический руководитель образовательной программы
Ю.А. Ландер

Москва, 2016

1 Область применения и нормативные ссылки

Настоящая программа учебной дисциплины устанавливает требования к образовательным результатам и результатам обучения студента и определяет содержание и виды учебных занятий и отчетности.
Программа предназначена для преподавателей, ведущих дисциплину Автоматическая обработка естественного языка, учебных ассистентов и студентов направления подготовки направления, обучающихся по образовательной программе Фундаментальная и прикладная лингвистика.
Программа учебной дисциплины разработана в соответствии с:

- Образовательным стандартом НИУ ВШЭ;
  Образовательной программой направления подготовки 45.03.03. Фундаментальная и прикладная лингвистика "Фундаментальная и прикладная лингвистика".
- Объединенным учебным планом университета по образовательной программе Фундаментальная и прикладная лингвистика, утвержденным в 2016 г.

2 Цели освоения дисциплины

Целями освоения дисциплины Автоматическая обработка естественного языка являются знакомство с основными проблемами в области компьютерной лингвистики, базовыми алгоритмами, математическими методами моделирования языковых феноменов, основными инструментами и технологиями в области автоматической обработки естественного языка, умение представлять в алгоритмическом виде процессы анализа и синтеза текста.

3 Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения дисциплины студент осваивает компетенции:

Компетенция	Код по ОС ВШЭ	Уровень формирования компетенции	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции	Форма контроля уровня сформированности компетенции
Способен планировать научно-исследовательскую деятельность, проводить самостоятельные исследования и получать новые научные результаты в области профессиональной деятельности	ПК-1	РБ, МЦ, СД	Дает определения основным понятиям автоматической обработки текста, воспроизводит базовые алгоритмы, используемы в автоматической обработке текста, использует основные пакеты морфологической обработки текста, демонстрирует знание базовых алгоритмов, владеет, использует современные методы тестирования качества, применяет современные подходы к решению задач в области компьютерной лингвистики, интерпретирует	- чтение специальной литературы -выполнение самостоятельных заданий - анализ полученных данных	Тест, проверка дз, сдача проекта

Компетенция	Код по ОС ВШЭ	Уровень формирования компетенции	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции	Форма контроля уровня сформированности компетенции
			результаты базовых алгоритмов][1]		
способен проводить формализацию лингвистических знаний, анализ и синтез лингвистических структур, квантитативный анализ лингвистических данных с использованием математических знаний и методов	ПК-2	РБ, МЦ, СД	дает определения основным формальным системам, используемым при разработке алгоритмов по морфологическому анализу текста, дает определения основным этапам синтаксического и морфологического анализа, воспроизводит различные алгоритмы синтаксического разбора, распознает основные типы ошибок автоматического и морфологического автоматического анализа, применяет квантитативные подходы к обработке текста и выделению ключевых слов в тексте		
способен участвовать в создании представительных текстовых массивов, корпусов текстов, корпусов звучащей речи, мультимодальных корпусов, лингвистических и социолингвистических баз данных и пользоваться этими ресурсами	ПК-11	РБ, СД	знает основные лингвистические ресурсы, владеет методами разметки корпусов и составлением частотных списков	практические занятия по созданию языковых ресурсов и лингвистических компонентов обзор и рефераты существующих разработок	Практические заняти, выполнение упражений, работа с ресурсами на практических занятиях
способен проектировать системы анализа и синтеза естественного языка, анализа и синтеза мультимодальных языковых систем, в том числе	ПК-12	РБ, СД	Знает наиболее известные доступные для свободного использования компоненты автоматического анализа, умеет использовать соответствующие модули в различных приложениях, в том числе синтаксические и морфологические парсеры, создавать модули первичной	Практические задания, выполнение проекта в группах	

[1] В шаблоне дан неполный перечень глаголов – подсказок. Возможно использование и других формулировок дескрипторов.

Компетенция	Код по ОС ВШЭ	Уровень формирования компетенции	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции	Форма контроля уровня сформированности компетенции
лингвистических компонентов интеллектуальных и информационных электронных систем			обработки текста		

4 Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к профессиональному блоку дисциплин «Компьютерная лингвистика».
Для специализаций Фундаментальная и компьютерная лингвистика профиля компьютерная лингвистика настоящая дисциплина является базовой.
Изучение данной дисциплины 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 дисциплин:

- автоматическая обработка естественного языка программы подготовки бакалавров на четвертом курсе
- информационный поиск и извлечение данных
- программирование для лингвистов
- подготовка и защита выпускной квалификационной работы

5 Тематический план учебной дисциплины

[Тематический план отражает содержание дисциплины (перечень разделов), структурированное по видам учебных занятий с указанием их объемов в соответствии с ОУП]

№	Название раздела	Всего часов	Аудиторные часы	Самостоя-

			Лекции	Семинары	Практические занятия	Другие виды работы[2]	тельная работа
1	Введение. Компьютерная лингвистика. Основные направления и задачи	8			4		4
2	Первичная обработка текста. Графематический анализ	18			8		10
3	Автоматический морфологический анализ	30			12		18
4	Автоматический синтаксический анализ	18			10		8
5	Контрольная	2			2		
6	Итого	76			36		40

6 Формы контроля знаний студентов

Тип контроля	Форма контроля				Параметры **
		2	3	4	
Текущий	Контрольная работа		*		письменная работа 120 минут
	Домашнее задание		*		Отчет по выполнению задания, требования к отчету регламентируются описанием требований к отчету по проекту
	Проект				
	Другие формы (указать)				
Итоговый	Экзамен		*		устный

** В графе Параметры указывается регламент (свод постоянных или временных правил, регулирующих внутреннюю организацию и формы деятельности) проведения контроля (заполняется для каждого контроля соответственно): формат работы (письменная, устная, тест, тест в компьютерной программе и другое), время, отведенное на аудиторные работы, количество дней проведения контроля, количество дней оценки результатов контроля (только для итогового контроля), объем письменных работ для домашних работ, сроки сдачи письменных работ (число), время на самостоятельную подготовку письменных работ и другая информация, носящая регламентирующий характер.]

7 Критерии оценки знаний, навыков

Данный курс в качестве текущего контроля предполагает выполнение практических заданий на семинарах и практических занятиях, а также одного домашнего практического задания,

[2] Указать другие виды аудиторной работы студентов, если они применяются при изучении данной дисциплины.

результаты которого представляются студентом в виде проекта на языке Python, а также технического отчета по результатам разработки системы сегментации текста и оценки системы морфологического анализа. Кроме того, на практических занятиях осуществляется постоянный текущий контроль в форме заданий, тестов, а также обсуждения текущих домашних упражнений.

При выполнении домашнего задания студент должен продемонстрировать знание основных проблем и принципов графематического анализа и первичной обработки текста, умение работать с основными корпусами текстов на русском языке, знание основных свободно-распространяемых систем морфологического анализа текста, умение запускать одну из систем, знание основных принципов анализа результатов морфологического анализа и принципов оценки качества морфологического анализа, умение анализировать результаты морфологического анализа.

При выполнении контрольной работы студенты должны продемонстрировать знание классификации основных задач компьютерной лингвистики, знание основных моделей и алгоритмов автоматического морфологического и синтаксического анализа текста, решать простые упражнения на применение базовых алгоритмов, быть в состоянии протестировать и оценивать работу отдельных модулей АОТ.

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

Домашнее задание по подготовке корпуса, разработке программы предварительного анализа текста, тестированию морфологических анализаторов являются групповыми проектами с индивидуальными заданиями, выполняемыми в рамках проектов.

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.

Сдача заданий и проверка происходят через LMS. Задания выполняются в формате тестов и проектов в LMS.

8 Содержание дисциплины

Раздел представляется в удобной форме (список, таблица). Изложение строится по разделам и темам. Содержание темы может распределяться по лекционным и практическим занятиям.

	практические занятия	Самостоятельная работа
Раздел1. Введение. Основные задачи компьютерной лингвистики		
1. Введение в компьютерную лингвистику. Задачи компьютерной лингвистики 2. Лингвистические системы. Этапы автоматической обработки текста	2 2	4
	4	4
Раздел 2. Первичная обработка текста.		

3. Предварительная обработка текста. Графематический анализ. Задачи, этапы и технологии первичной обработки текста. Проблемы токенизации. Анализ различных подходов к токенизации и делению на предложение. 4. Поиск, индексация, оценка качества. Частотный анализ лексики и ключевые слова 5. Различные подходы к токенизации. Методы разбиения на предложения 6. Оценка качества токенизаторов (коллоквиум, отчет по проекту, оценка качества)	2 2 2 2	2 8
	8	10
Раздел 3. Автоматический морфологический анализ		
7. Автоматический морфологический анализ. Введение 8. Морфологический анализ: организация данных 9. Конечные автоматы и конечные преобразователи в морфологическом анализе 10. Методы дизамбигуации (машинное обучение: rule-induction, HMM). 11. **Коллоквиум.** Оценка качества морфологических анализаторов	2 4 2 4	4 6 8
	12	18
Раздел 4. Автоматический синтаксический анализ		
16. Формализмы и методы автоматического синтаксического анализа 17. Алгоритмы автоматического синтаксического анализа. Анализ работы систем 18. Синтаксический анализ в терминах деревьев зависимостей	2 2 2	4 4
	10	8
Контрольная работа	2	

.

Раздел1. Введение. Основные задачи компьютерной лингвистики

8.1. Тема 1. Введение в автоматическую обработку текста

Практическое занятие 1

Задание

Рассмотреть https://visl.sdu.dk/; http://siberian-lang.srcc.msu.ru/ru/sintaksichyeskiy_proyekt_RFFI

Ответить на вопросы:

- какие типы лингвистических электронных ресурсов необходимы в лингвистических исследованиях
- какие типы лексикографических ресурсов используются в изучении и исследовании языка

Домашнее задание:

Чтение литературы:

[J&M], статья из списка на выбор: выписать три термина из статьи (русский и английский вариант), привести его краткое определение, привести пример

8.2. Тема 2. Лингвистические системы. Этапы автоматической обработки текста

Практическое занятие 2

8.3. Тест по прочитанной литературе

8.4. Рассмотреть поисковик Яндекс (yandex.ru) и Яндекс-новости (news.yandex.ru).

На основе тестирования систем

(а) сформулировать основные задачи обработки контента;

(б) перечислить основные лингвистические задачи, которые решают системы.

Литература

1. Прикладная лингвистика. // Статья в энциклопедии «Фонд знаний «Ломоносов»». URL: http://www.lomonosov-fund.ru/enc/ru/encyclopedia:01206:article
2. [J&M] - Introduction // Daniel Jurafsky & James H. Martin. Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition. Second edition. 2009. http://stp.lingfil.uu.se/~santinim/ml/2014/JurafskyMartinSpeechAndLanguageProcessing2ed_draft%202007.pdf
3. Большакова и др. (2011). Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : Часть 2, глава 5-6.

Дополнительная литература

Генерация текстов

Соколова Е.Г., Болдасов М. Автоматическая генерация текстов на ЕЯ (портрет направления) http://www.dialog-21.ru/Archive/2004/Sokolova.htm
Ehud Reiter. Has a Consensus NL Generation Structure Appeared, and is it Psycholinguistically Plausible? — 7th International Generation Workshop (Kennebunkport, Maine). URL: http://www.aclweb.org/anthology/W/W94/W94-0319.pdf

Машинный перевод

Лекция Л.Л.Иомдина "Машинный перевод: успехи, неудачи, надежды". Лекторий Политехнического музея. Видео. Доступно с URL http://rutube.ru/video/828268c50a49b876a6f4676b839fa745/ дата обращения 20.01.2014)
Brown P. F. et al. The mathematics of statistical machine translation: Parameter estimation //Computational linguistics. – 1993. – Т. 19. – №. 2. – С. 263-311. http://acl.ldc.upenn.edu/J/J93/J93-2003.pdf

Text mining (классификация, кластеризация, реферирование). На примере анализа новостного потока

Кондратьев М. Е. Анализ методов кластеризации новостного потока //Тр. Восьмой Всерос. науч. конф.«Электронные библиотеки: перспективные методы и технологии, электронные коллекции»(RCDL'2006).—Ярославль. – 2006. – С. 108-114.rcdl.ru/doc/2006/paper_92_v1.pdf

Распознавание речи

Speech recognition. http://en.wikipedia.org/wiki/Speech_recognition
В. Н. Сорокин, *Синтез речи*. М., 1992.,
D. Childers, *Speech Processing*, John Wiley and Sons, Inc., все издания, начиная с 1999

Диалоговые системы. Говорящие роботы

Filipe P. P., Morgado L., Mamede N. J. An Adaptive Domain Knowledge Manager for Dialogue Systems //ICEIS (5). – 2007. – С. 45-52. http://www.inesc-id.pt/pt/indicadores/Ficheiros/3983.pdf
Bermúdez M. G., Vila M. G. Dialogue Management for multilingual communication through different channels.

Извлечение информации из текста: извлечение именованных сущностей, отношений и фактов

Nadeau D. and Sekine S. A survey of named entity recognition and classification, Linguisticae Investigationes, Amsterdam, Netherlands: John Benjamins Publishing Company, 1: Vol. 30. pp. 3-26.

Brykina M., Toldova S.Yu., Faynveyts A. V. Dictionary-based ambiguity resolution in Russian named-entities recognition. A case study. P. 163-177. Компьютерная лингвистика и интеллектуальные технологии: по материалам международной конференции "Диалог" 2013. Вып. 12(19). М.: РГГУ, 2013. URL: http://www.dialog-21.ru/digests/dialog2013/materials/pdf/BrykinaMM.pdf

Раздел 2. Первичная обработка текста

8.3. Тема 3. Предварительная обработка текста. Графематический анализ

Сегментация текста. Создание собственного корпуса текстов. Очистка текста. Графематический анализ. Признаки токенов. Адрес токена (смещение). Обработка

Практическое занятие 3.

Задание

Каждая группа получает фрагмент текста на русском языке.
Разбить текст на токены. Выписать случаи, которые вызвали вопросы. Предложить необходимые компоненты сегментации текста (например, отдельный модуль для обработки адресов веб-страниц). По результатам анализа токенизации и сегментации на предложения. Сравнить результаты разбиения текста в группе

Домашние упражнения

1) Проанализировать текст с точки зрения задач препроцессинга. Сформулировать правила обработки буквенно-цифровых комплексов, сокращений, слов с дефисом.
2) Разбить на токены предложение на китайском языке, предложить алгоритм разбиения
3) Выбрать одну из статей. Подготовить по ней презентацию, краткое резюме, выбрать три-пять терминов из статьи, привести переводные эквиваленты и определения.

Статьи для чтения:

[1] M. Aulbach, S. Evert, and B. Schrader, "Requirements for and design of a flexible tokenization system," 2006.
[2] B. Habert, G. Adda, P. B. De Mar, S. Ferrari, O. Ferret, G. Illouz, P. Paroubek, and F.-O. Cedex, "Towards Tokenization Evaluation."
[3] L. Karttunen, J.-P. Chanod, G. Grefenstette, and A. Schille, "Regular expressions for language engineering," Nat. Lang. Eng., vol. 2, no. 4, pp. 305–328, Dec. 1996.
[4] Y. Liu and E. Shriberg, "Comparing Evaluation Metrics for Sentence Boundary Detection," Acoust. Speech Signal Process. 2007. ICASSP 2007. IEEE Int. Conf., vol. 4, pp. 451–458, 2007.
[5] M. Stevenson and R. Gaizauskas, "Experiments on sentence boundary detection," Proc. sixth Conf. Appl. Nat. Lang. Process. -, pp. 84–89, 2000.
[6] С. В. Алексеева, Д. В. Грановский, Н. А. Остапук, М. Е. Степанова, and А. В. Суриков, "Сегментация текста в проекте « Открытый корпус » Text segmentation in opencorpora project."
[8] В. В. Бочаров and Д. В. Грановский, "Вероятностная модель токенизации в проекте Открытый корпус," Новые информационные технологии в автоматизированных системах материалы пятнадцатого научно-практического семинара, pp. 1–8, 2012.
[9] "AOT :: Технологии :: Графематика: программный интерфейс." [Online]. Available: http://aot.ru/docs/graphan.html. [Accessed: 24-Sep-2014].
[10] G. Laboreiro, L. Sarmento, J. Teixeira, and E. Oliveira, "Tokenizing micro-blogging messages using a text classification approach," Proc. fourth Work. Anal. noisy unstructured text data - '10, p. 81, 2010.

8.4. Тема 4. Поиск, индексация, частотное распределение языковых элементов в тексте, оценка качества

Индекс. Обратный индекс. Конкорданс

Частотное распределение лексики в языке. Закон Ципфа. Доля hapax legomena. Скорость роста словаря. Меры лексического разнообразия и их применимость.

Распределение лексики в текстах коллекции. Взвешенная частотность. TF-IDF. Прочие меры лексической дисперсии. Мера отклонения пропорций DP и DPnorm.

Понятие релевантности документа, точности и полноты информационного поиска

Практическое занятие 4.

Построение индекса
Построение частотного списка. ПО для построения частотных списков лексики: AntConc (http://www.laurenceanthony.net/software/antconc/). Лексический состав вершины, середины, хвоста частотного списка. Контрастный анализ корпусов на основании частотных списков. Вычисление tf.idf с использованием корпуса со снятой омонимии Национального корпуса русского языка (НКРЯ).

Домашние упражнения

- Выбрать коллекцию текстов в подкорпусе со снятой омонимией НКРЯ. Выбрать 8 слов (2 «тематических», 2 общеупотребительных из средней части частотного списка, 2

высокочастотных общеупотребительных, 2 редких). Используя статистику НКРЯ, вычислить tf.idf, упорядочить слова по убыванию tf.idf. Результат проанализировать.

Чтение

[Маннинг&Рагхаван& Шютце]

8.5. **Задачи, этапы и технологии первичной обработки текста. Проблемы токенизации. Анализ различных подходов к токенизации и делению на предложения**

Сегментация на предложения. Вектор признаков.

Практическое занятие 5

Доклад по одной из выбранных статей.

8.6. Оценка качества токенизаторов (коллоквиум, отчет по проекту, оценка качества)

Практическое занятие 6

Соревнование токенизаторов и сплиттеров. Сравнение качества. Доклады по проектам.

Литература к разделу

Маннинг К. Д., Рагхаван П., Шютце Х. Введение в информационный поиск.: Пер. с англ. – М.: ООО «Вильямс», 2011. 6.2 Частота термина и взвешивание; 6.4 Варианты функций tf-idf.

M. Baroni Distributions in text. In Anke Lüdeling and Merja Kytö (eds.), *Corpus Linguistics: An International Handbook*. Berlin: Mouton de Gruyter, 2008.

Сегалович "Как работают поисковые системы" http://download.yandex.ru/company/iworld-3.pdf

Ch 3. Sec. 3.9 Jurafsky and Martin, Speech and Language Processing, 2nd Edition (2009)

Christopher Potts. Sentiment Symposium Tutorial: Tokenizing. http://sentiment.christopherpotts.net/tokenizing.html

Автоматическая обработка текста. Графематика. http://www.aot.ru/docs/graphan.html

Урюпина, О. Автоматическое разбиение текста на предложения для русского языка // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2008 г.). Вып. 7 (14). – М.: РГГУ, 2008, с. 539–544. **http://www.dialog-21.ru/digests/dialog2008/materials/html/83.htm**

Раздел 3. Автоматический морфологический анализ

8.7. Тема 7. Автоматический морфологический анализ. Введение

Основные задачи и этапы морфологического анализа. Нормализация (стемминг и лемматизация); грамматическое аннотирование; морфологический парсинг; дизамбигуация; морфологический анализ незнакомых слов

Праткическое занятие 7

Обсуждение основных задач и этапов морфологического анализа в зависимости от сферы его применения. Обсуждение основных явлений морфологии, влияющих на качество морфологического анализа: морфонологические явления, виды морфологической омонимии; различные типы незнакомых слов; многословные функциональные единицы (союзы, предлоги, аналитические формы).

8.8. Тема 8. Морфологический анализ: организация данных

Праткическое занятие 8

Разбор различных способов представления данных для морфологического анализа:

(а) существительные с беглыми гласными в русском языке: сколько парадигм необходимо для их описания

(б) описание морфонологических чередований

(в) порядок морфем

8.9. Конечные автоматы и конечные преобразователи в морфологическом анализе

Практическое занятие 8

Построение конечных автоматов, распознающих словоформы, в которых произошли морфонологические чередования при присоединении словоизменительных морфем

Построение конечных преобразователей для описания всех форм глаголов в английском

Построение конечного преобразователя для описания замены –y (try -> tries) при в английском

Домашние упражнения

Предложить два различных формальных описания глагольного словоизменения (разные способы разбиения на основу и окончание, + способы описания морфонологических чередований)

Построить конечный автомат, допускающий только правильные словоформы (словоформы, при построении которых были правильно применены правила сингармонизма / морфонологических чередований)

Построить конечный преобразователь для описания сингармонизма в одном из тюркских языков

8.10. Методы дизамбигуации (машинное обучение: rule-induction, HMM).

Методы снятия морфологической омонимии. Извлечение правил. Понятие марковской модели. Понятие скрытой марковской модели. Основные допущения при применении скрытой марковской модели к частеречной обработке

Практическое занятие 10.

Знакомство со скрытыми марковскими моделями и их применением к дизамбигуации. Вычисление условной вероятности частеречных тегов, вычисление лексической вероятности тегов.

8.11. Оценка качества морфологических анализаторов

Инструменты морфологического анализа для русского языка. Mystem. AOT. pymorphy. TreeTagger. TnT. FreeLing

Практическое занятие 11.

Защита проекта оценки качества морфологического анализатора (вторая часть домашнего задания). Сдача отчета по тестированию одного из морфологических анализаторов с дизамбигуацией: TreeTagger, TnT, FreeLing

Проект "Тестирование морфологического анализатора"

Запустить один из морфологических анализаторов для русского языка из списка. Провести тестирование системы: функциональное тестирование; вычислить accuracy.

Литература к разделу

Jurafsky D., James H. Speech and language processing an introduction to natural language processing, computational linguistics, and speech. – 2000. Ch. 2. Regular Expressions and Automata. http://people.mokk.bme.hu/~kornai/termeszetes/3.pdf

Jurafsky D., James H. Speech and language processing an introduction to natural language processing, computational linguistics, and speech. – 2000. Ch. 3. Words and transducers http://people.mokk.bme.hu/~kornai/termeszetes/3.pdf

Коваль С.А. Лингвистические проблемы компьютерной морфологии. СПб., 2005.

Дополнительная литература

Ю.Г. Зеленков, И.В. Сегалович, В. А. Т. (2005). Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов. Retrieved from http://www.dialog-21.ru/Archive/2005/Zelenkov Segalovich/Zelenkov_Segalovich.htm

Сокирко, А. В., & Толдова, С. Ю. (2004). Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка (скрытая модель Маркова и синтаксический анализатор именных групп), (http://www.aot.ru/docs/RusCorporaHMM.htm).

Ляшевская О. Н., Астафьева И., Бонч-Осмоловская А., Гарейшина А., Гришина Ю., Дьячков В., Ионов М., Королева А., Кудринский М., Литягина А., Лучина Е., Сидорова Е., Толдова С. "Оценка методов автоматического анализа текста: морфологические парсеры русского языка". // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26-30 мая 2010 г.). Вып. 9 (16). М.: Изд-во РГГУ, 2010.

Используемые для выполнения заданий ресурсы и системы

Морфологические частеречные разметчики

- FreeLing http://nlp.lsi.upc.edu/freeling/
- Stanford Log-linear Part-Of-Speech Tagger http://nlp.stanford.edu/software/tagger.shtml
- TreeTagger http://corpus.leeds.ac.uk/mocky/
- FreeLing: http://nlp.lsi.upc.edu/freeling/index.php?option=com_content&task=view&id=18&Itemid=47

Ресурсы и библиотеки для реализации FSA и FST

1) PC-Kimmo - http://www-01.sil.org/pckimmo/
2) foma - https://code.google.com/p/foma/
3) Xerox Finite-State Tool (Lauri Karttunen, Tamás Gaál, and André Kempe) - http://www.cis.upenn.edu/~cis639/docs/xfst.html

Справочные материалы по FST

FST Morphology - http://web.stanford.edu/~laurik/fsmbook/home.html

Раздел 4. Автоматический синтаксический анализ

8.16. Формализмы и методы автоматического синтаксического анализа

Синтаксические отношения. Синтаксическая омонимия. Непосредственные составляющие. Зависимости.

Практическое занятие 16.

Синтаксический анализ предложения. Случаи синтаксической омонимии, синтаксические нули. Формальное представление синтаксической структуры предложения.

8.17. Алгоритмы автоматического синтаксического анализа. Анализ работы систем

Контекстно-свободная грамматика. Нормальная форма Хоского. Алгоритм Кока-Янгера-Касами. Лексикализованные грамматики. Вероятностные контекстно-свободные грамматики

Практическое занятие 17.

Правила перевода контекстно-свободной грамматики в нормальную форму Хомского. Анализ предложения с использованием алгоритма Кока-Янгера-Касами.

8.18. Синтаксический анализ в терминах деревьев зависимостей

Зависимости. Критерий эндоцентричности. Критерий морфосинтаксического локуса. Непроективность.

Практическое занятие 18.

Разбор алгоритмов синтаксического анализа в терминах зависимостей. Работа с синтаксически-размеченными корпусами. Анализ ошибок автоматического синтаксического парсинга

Литература к разделу

Jurafsky D., James H. Chapter 13. Parsing with Context-free Grammar. Speech and language processing an introduction to natural language processing, computational linguistics, and speech. – 2009. The 2nd edition (Chapter 9. Edition - 2000).

Апресян Ю. Д. и др. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы //Национальный корпус русского языка. – 2003. – Т. 2005. – С. 193-214. URL: http://corpora.phil.spbu.ru/Works2008/Boguslavsky1_56_74.pdf (дата обращения: 27.01.2015)
Joakim Nivre, Sandra Kubler. Dependency Parsing. Tutorial at COLING-ACL, Sydney 2006. (презентация) http://stp.lingfil.uu.se/~nivre/docs/ACLslides.pdf

Дополнительная литература

Анисимович К. В. и др. Синтаксический и семантический парсер, основанный на лингвистических технологиях ABBYY Compreno. // В кн. Компьютерная лингвистика и интеллектуальные технологии. (дата обращения: 27.01.2015)
По материалам ежегодной Международной конференции "Диалог" (2012). Том 2. Доклады специальных секций – (URL: http://www.dialog-21.ru/digests/dialog2012/materials/pdf/anisimovich.pdf (дата обращения: 27.01.2015))
Иомдин Л.Л. и др. Синтаксический анализатор системы ЭТАП: современное состояние. // В кн. Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции "Диалог" (2012). Том 2. Доклады специальных секций – (URL: dialog2012/materials/pdf/Iomdin.pdf (дата обращения: 27.01.2015))
Антонова А. А., Мисюрев А. В. Анализатор русского языка syntautom для соревнования синтаксических парсеров (диАлог-2012). // В кн. Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции "Диалог" (2012). Том 2. Доклады специальных секций. (http://www.dialog-21.ru/digests/dialog2012/materials/pdf/98.pdf (дата обращения: 27.01.2015))

Корпуса для анализа:

- СинТагРус http://ruscorpora.ru/search-syntax.html (дата обращения: 27.01.2015),
- Тестовый корпус с параллельной синтаксической разметкой http://otipl.philol.msu.ru/~soiza/testsynt/, (дата обращения: 27.01.2015),
- Rus-Treebank http://otipl.philol.msu.ru/~soiza/rtb/res01/rtb.php (дата обращения: 27.01.2015)

9 Образовательные технологии

[Основой для знакомства с методологией анализа текстов и овладения практическими навыками автоматического анализа текста с помощью программных средств в рамках курса служит работа над учебным проектом, состоящим из двух частей, по анализу текстовой коллекции. В начале курса студентам предлагается на выбор несколько жанров. На протяжении курса студенты работают над сбором и анализом коллекции текстов в малых группах (2-3 человека). В задачи проекта входит создание собственного токенизатора и сплиттера. А также морфологический анализ текста. Тестирование одного из морфологичемских тегеров. По результатам анализа группы пишут отчет по проекту, который состоит из (а) аналитиечской записки, содержащей краткое изложение существующих подходов к задаче, стандартных проблем, описанных в литературе по данному вопросу; (б) краткие характеристики системы; (в) результаты тестирования системы с описанием сложных случаев и анализом ошибок.
В качестве простых упражнений рекомендуется также упражнения на пошаговое применение обсуждаемых в курсе алгоритмов вручную.
Для освоения материала на практических занятиях используются задания, предназначенные как для индивидуального решения задач по обработке текста, так и для коллективного обсуждения стратегии решения той или иной задачи. Проводится обсуждение отдельных методов компьютерной лингвистики в форме мини-докладов студентов по материалам проведенного анализа выбранной текстовой коллекции. Особое внимание уделяется организации самостоятельной работы студентов с программным обеспечением, рассматриваемым в рамках курса.

10 Оценочные средства для текущего контроля и аттестации студента

10.1 Оценочные средства для оценки качества освоения дисциплины в ходе текущего контроля

Примерный перечень вопросов к различным формам текущего контроля.

1. **Задачи и направления компьютерной лингвистики**
   3 направления компьютерной лингвистики. Основные задачи в рамках каждого направления. Типы систем и задач
2. **Задачи автоматического анализа текста**
   Проверка правописания, грамматики и стиля. Распознавание текстов. Распознавание и синтез речи. Машинный перевод текста и речи (классика NLP). Информационный поиск. Реферирование. Классификация (кластеризация и рубрикация) текстов, установление сходства текстов (плагиат и т.п.). Автофильтрация. Вопросно-ответные и диалоговые системы. Системы извлечения знаний (извлечение именованных сущностей, извлечение отношений и фактов, извлечение тональности/мнений). Примеры систем
3. **Типы лингвистических данных**
   Типы электронных ресурсов, представляющих лингвистические данные: лексикографические ресурсы; корпуса. Примеры
4. **Корпуса текстов. Основные понятия корпусной лингвистики**
   Корпус текстов: определение, решаемые задачи. Национальный корпус текстов. Типы корпусов. Понятие сбалансированного vs. мониторного корпуса. Требования к корпусам: репрезентативность, полнота, структурированность. Единицы хранения. Типы аннотации корпусов. Назначение корпусов. Использование корпусов при разработке систем автоматического анализа текстов.
5. **Основные этапы автоматической обработки текста**
   Задачи каждого из этапов. Проблемы лингвистической обработки текста на разных этапах: неоднозначность, несимметричность, избыточность, конвенциональность, эллиптичность и т.п.
6. **Информационный поиск**
   Полнотекстовый поиск vs. индекс. Термин vs. лексема. Матрица термин-документ. Инвертированный индекс. Токен vs. лемма. Частотные характеристики элементов текста (лемм, n-грам).
7. **Этапы предварительной обработки текста.**
   Сбор корпуса текстов. Проблема унификации текстов в корпусе (кодировка, «многозначность» служебных символов, нетекстовые элементы и т.п.). Стандарты представления текстов (например, TEI). Предварительная обработка текста. Задачи и проблемы токенизации. Особенности токенизации для разных типов и жанров текстов. Сегментация текстов на предложения. Методы сегментации.
8. **Автоматический морфологический анализ**
   Типы задач автоматического морфологического анализа: нормализация (стемминг и лемматизация); частеречная/лекско-грамматическая аннотация; полный морфологический анализ; дизамбигуация
9. **Этапы автоматического морфологического анализа:**
   Токенизация, лемматизация, первичная частеречная аннотация, дизаимбигуация, идиоматизация, постобработка. Лингвистические проблемы, решаемые на различных этапах автоматического морфологического анализа.
10. **Подходы к представлению данных в автоматическом морфологическом анализе.**
    Словарь vs. правила. Контекст словоформы vs. внутренняя структура словоформы. Словарь словоформ (информация, используемая в словаре словоформ, для каких методов анализа используется чаще). Понятие основы и окончания при автоматическом морфологическом анализе, понятие парадигмы (псевдоосновы и псевдоокончания). Правила порождения словоформ (правила перехода от глубинного представления словоформы к поверхностному (учет морфонологических чередований и т.п.)

11. **Технологии реализации морфологического анализа**
    Основные технологии: конечные автоматы; конечные преобразователи; регулярные выражения. Моделирование правил перехода от лексического уровня представления словоформы к промежуточному. Моделирование правил перехода от промежуточного уровня к поверхностному.
12. **Методы частеречной разметки. Основные методы снятия морфологической омонимии**
    Классификация методов лексико-грамматической (частеречной разметки). Организация грамматических аннотаций. Методы, основанные на правилах. Метод извлечения правил (трансформационный метод) Э.Брилла. Метод, основанный на скрытых марковских моделях.
13. **Основания оценки качества морфологического разбора**
    Оценка общего качества работы модуля автоматической обработки текста. Разработка золотого стандарта. Метрики качества морфологического анализа: полнота, точность, степень омонимичности. Классификация ошибок. Проблемы тестирования систем морфологического анализа. Функциональное тестирование.
14. **Автоматический синтаксический анализ. Задачи. Способы представления синтаксической структуры**
    Приложения, в которых используется автоматический синтаксический анализ. Задачи автоматического синтаксического анализа. Основные единицы анализа. Формальные системы представления синтаксической структуры предложения
15. **Автоматический синтаксический анализ. Лингвистические проблемы**
    Синтаксическая омонимия. Наиболее проблемные случаи для разрешения синтаксической омонимии при автоматическом синтаксическом анализе. Синтаксические нули. Способы решения. Нестандартный порядок слов (перемещения, разрывы составляющих). Непроективные структуры. Дальние связи. Эллипсис и нули.
16. **Контекстно-свободные грамматики.**
    Особенности представления предложений в терминах деревьев непосредственных составляющих при автоматическом синтаксическом анализе. Проблемы: перемещения и разрывы составляющих, дальние связи. Способы решения проблем: «размножение» правил с учетом субкатеогризации; вероятностные контекстно-свободные грамматики; лексикализованные вероятностные контекстно-свободные грамматики
17. **Алгоритмы автоматического синтаксического анализа в терминах непосредственных составляющих**
    Основные алгоритмы реализации автоматического синтаксического анализа в терминах непосредственных составляющих. Метод нисходящего спуска. Восходящий анализ. Алгоритм Кока-Янгера-Касами (CKY Parsing).
18. **Автоматический анализ в терминах деревьев зависимостей.**
    Основные подходы. Проблемы. Случаи неоднозначного определения вершины и зависимого. Непроективность. Нестандартные решения при выборе направления связей
19. **Методы улучшения качества работы систем синтаксического анализа.**
    Использование лексической и онтологической информации в системах автоматического морфологического анализа. Использование информации о частотных свойствах единиц синтаксического анализа (какие частотные характеристики используются, часта каких единиц учитывается).
20. **Базовые алгоритмы анализа в терминах зависимостей**
    Переход к деревьям НС. Грамматика ограничений. Алгоритм Нивра.
21. **Кратко опишите одну из систем автоматического синтаксического анализа**
    Основное предназначение. Внутренняя организация. Используемые модули. Используемые базовые алгоритмы. Подходы к решению проблемных случаев (непроективность, нули и т.п.). Методы «борьбы» с синтаксической неоднозначностью.
22. **Частичный синтаксический анализ**

Задачи, в которых используется неполный синтаксический анализ (shallow parsing). Выделение связанных фрагментов предложения (chunking). Основные задачи и технологии.

11 Практические, аналогичные тем, которые были в ДЗ:

1. Для некоторого морфонологического процесса сформулировать правило; построить конечный автомат, распознающий словоформы, получившиеся в результате действия данного правила
2. Предложить правила построения некоторого класса словоформ: построить конечный преобразователь, задающий переход от лексического уровня к промежуточному или от промежуточного к поверхностному
3. Для некоторой цепочки словоформ с морфологической аннотацией сравнить 2 варианта последовательности морфологических тегов, найти более вероятную последовательность
4. Дано множество предложений: построить для них деревья непосредственных составляющих; построить кс-грамматику, которая порождает все данные деревья; применить ее к новому предложению
5. Дана кс-грамматика и предложение: перевести ее в нормальную форму Хомского и расписать шаги алгоритма CYK
6. Дана предложение; построить дерево зависимостей и описать места, проблемные для анализа в терминах деревьев зависимостей
7. Даны несколько предложений в корпусе с параллельной синтаксической разметкой; проанализировать расхождения в ответах систем; определить, какие расхождения объясняются разными теоретическими решениями; какие - являются ошибками системы

Бонусное задание:

8. Распишите отдельные шаги/условия применения одного из базовых алгоритмов для построения деревьев зависимостей для предложения

Дополнительные задания

9. Найти в тексте сложные случаи для предварительной обработки, предложить правила решения
10. Найти в тексте сложные случаи морфологического анализа текста для разных этапов морфологического анализа
11. Найти в предложении случаи, сложные для решения в системах автоматического синтаксического анализа в терминах непосредственных составляющих

11.1 Примеры заданий промежуточной аттестации

Пример контрольной работы

Часть 1. Без компьютера

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.*

12 Порядок формирования оценок по дисциплине

Преподаватель оценивает работу студентов на практических занятиях: активность студентов в дискуссиях, правильность выполнения заданий (на основании отчетов по практическим заданиям), выполнение мини-тестов по теории. Оценки за работу на практических занятиях преподаватель выставляет в рабочую ведомость. Оценка по 10-ти балльной шкале за работу на практических занятиях определяется перед промежуточным или итоговым контролем - $O_{аудиторная}$.

Преподаватель оценивает самостоятельную работу студентов: правильность выполнения домашних работ, задания для которых выдаются на практических занятиях (*имеются ввиду домашние работы, которые не включаются в ОУП, это не форма текущего контроля "Домашнее задание"*), полнота освещения темы, которую студент готовит для выступления с докладом на занятии-дискуссии. Оценки за самостоятельную работу студента преподаватель выставляет в рабочую ведомость. Оценка по 10-ти балльной шкале за самостоятельную работу определяется перед промежуточным или завершающим контролем - $O_{сам}$.
Накопленная оценка по дисциплине рассчитывается по формуле:
$O_{накопленная}$= 0,45* $O_{текущий}$ +0,1* $O_{ауд}$ + 0,45* $O_{сам.работа}$.]

Накопленная оценка за текущий контроль учитывает результаты студента по текущему контролю следующим образом:

$$O_{текущий} = 0.4 \cdot O_{контр} + 0.6 \cdot O_{дз};$$

Способ округления накопленной оценки текущего контроля: – арифметический.
В диплом выставляется результирующая оценка по учебной дисциплине.
$O_{результ} = 0,71* O_{накопл} + 0,29 *\cdot O_{экз}$]

Способ округления результирующей оценки по учебной дисциплине: [арифметический].

Примеры расчета оценки приведены в Приложении 1.

13 Учебно-методическое и информационное обеспечение дисциплины

13.1 Базовый учебник

[*Jurafsky, D., Martin, J. H.* (2000) Speech and language processing. NJ: Prentice Hall, 2000.. http://www.deepsky.com/~merovech/voynich/voynich_manchu_reference_materials/PDFs/jurafsky_martin.pdf

13.2 Основная литература

[*Большакова и др. (2011). Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с. http://clschool.miem.edu.ru/uploads/swfupload/files/011a69a6f0c3a9c6291d6d375f12aa27e349cb67.pdf*
].
Jurafsky, Daniel, and James H. Martin. (2009). Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics . 2nd edition. Prentice-Hall.
Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце Введение в информационный поиск, М. Вильямс, 201

13.3 Дополнительная литература

- Автоматическая обработка текста. Графематика. http://www.aot.ru/docs/graphan.html
- Анисимович К. В. и др. Синтаксический и семантический парсер, основанный на лингвистических технологиях ABBYY Compreno. // В кн. Компьютерная лингвистика и интеллектуальные технологии. (дата обращения: 27.01.2015)
- Антонова А. А., Мисюрев А. В. Анализатор русского языка syntautom для соревнования синтаксических парсеров (диАлог-2012). // В кн. Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции "Диалог" (2012). Том 2. Доклады специальных секций. (http://www.dialog-21.ru/digests/dialog2012/materials/pdf/98.pdf (дата обращения: 27.01.2015))

- Апресян Ю. Д. и др. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы //Национальный корпус русского языка. – 2003. – Т. 2005. – С. 193-214. URL: http://corpora.phil.spbu.ru/Works2008/Boguslavsky1_56_74.pdf (дата обращения: 27.01.2015)
- Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. Лингвистический процессор для сложных информационных систем
- Екатерина Протопопова, Виктор Бочаров. Автоматическое извлечение правил для снятия морфологической неоднозначности. // В кн.: Доклады всероссийской научной конференции АИСТ'2013 / Отв. ред.: Е. Л. Черняк; науч. ред.:Д. И. Игнатов, М. Ю. Хачай, О. Баринова. М. : Национальный открытый университет «ИНТУИТ», 2013. С. 184-95. URL: http://aistconf.org/stuff/aist2013/submissions/aist2013_submission_28.pdf
- Зеленков Ю.Г., Сегалович И.В., Титов В.А. Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов. // Компьютерная лингвистика и интеллектуальные технологии. Труды международного семинара Диалог'2005. – М., 2005. URL: http://www.dialog-21.ru/Archive/2005/Zelenkov%20Segalovich/Zelenkov_Segalovich.htm (http://download.yandex.ru/company/Zelenkov_Segalovich.pdf),
- Иомдин Л.Л. и др. Синтаксический анализатор системы ЭТАП: современное состояние. // В кн. Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции "Диалог" (2012). Том 2. Доклады специальных секций – (URL: dialog2012/materials/pdf/Iomdin.pdf (дата обращения: 27.01.2015))
- Искусственный интеллект: Справочник: Кн.1: Системы общения и экспертные системы. - М.: Радио и связь, 1990.
- Коваль С. А. Лингвистические проблемы компьютерной морфологии. - СПб.: Изд-во С.-Петерб. ун-та, 2005. - 151 с.
- Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце. Введение в информационный поиск, М. 2011, Вильямс Гл. 1 пп.1.1.-1.2., гл.2. пп. 2.1-2.2.
- Лакомкин Е. Д., Рыжова Д. А., Пузыревский И. Анализ статистических алгоритмов снятия морфологической омонимии в русском языке // В кн.: Доклады всероссийской научной конференции АИСТ'2013 / Отв. ред.: Е. Л. Черняк; науч. ред.:Д. И. Игнатов, М. Ю. Хачай, О. Баринова. М. : Национальный открытый университет «ИНТУИТ», 2013. С. 184-195. URL: http://aistconf.org/stuff/aist2013/submissions/aist2013_submission_33.pdf
- По материалам ежегодной Международной конференции "Диалог" (2012). Том 2. Доклады специальных секций – (URL: http://www.dialog-21.ru/digests/dialog2012/materials/pdf/anisimovich.pdf (дата обращения: 27.01.2015))
- Сокирко А.В., Толдова С.Ю. Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка (скрытая модель Маркова и синтаксический анализатор именных групп). URL: http://www.aot.ru/docs/RusCorporaHMM.htm,
- Структурная и прикладная лингвистика. Под ред. А. С. Герда. Вып. 1. Л., 1978. — Вып. 7. СПб., 2008
- Урюпина. О. Автоматическое разбиение текста на предложения для русского языка. Результат: три вопроса к тексту http://www.dialog-21.ru/digests/dialog2008/materials/html/83.htmСправочники, словари, энциклопедии

13.4 Программные средства

Для успешного освоения дисциплины, студент использует следующие программные средства:

- Программа обработки текстовых массивов (составление конкордансов и частотных словарей) AnrConc. http://www.laurenceanthony.net/software/antconc/
- mystem. Морфологический анализатор для русского языка. http://company.yandex.ru/technologies/mystem/
- Python 3
- FreeLing http://nlp.lsi.upc.edu/freeling/
- Stanford Log-linear Part-Of-Speech Tagger http://nlp.stanford.edu/software/tagger.shtml
- TreeTagger http://corpus.leeds.ac.uk/mocky/
- FreeLing: http://nlp.lsi.upc.edu/freeling/index.php?option=com_content&task=view&id=18&Itemid=47

13.5 Дистанционная поддержка дисциплины

Материалы курса представлены в LMS
Для освоения программы используются электронные ресурсы:

http://text0.mib.man.ac.uk:8080/scottpiao/sent_detector
http://beta.visl.sdu.dk/visl/en/parsing/automatic/parse.php
http://ruscorpora.ru/
http://www.corpus-i.compling.net/res01/rtb.php
http://www.connexor.fi/

- СинТагРус http://ruscorpora.ru/search-syntax.html (дата обращения: 27.01.2015),
- Тестовый корпус с параллельной синтаксической разметкой http://otipl.philol.msu.ru/~soiza/testsynt/, (дата обращения: 27.01.2015),
- Rus-Treebank http://otipl.philol.msu.ru/~soiza/rtb/res01/rtb.php (дата обращения: 27.01.2015)

14 Материально-техническое обеспечение дисциплины

Практические занятия проводятся с использованием мультимедийного проектора в компьютерных классах [

Приложение 1

Примеры подсчета оценки за дисциплину в различных случаях

Накопленная оценка за текущий контроль учитывает результаты студента по текущему контролю следующим образом:

$$О_{накопленная} = k_1 * О_{текущий} + k_2 * О_{ауд} + k_3 * О_{сам.работа}$$

где $О_{текущий}$ рассчитывается как взвешенная сумма всех форм текущего контроля, предусмотренных в ОУП

$$О_{текущий} = n_1 \cdot О_{эссе} + n_2 \cdot О_{к/р} + n_3 \cdot О_{реф} + n_4 \cdot О_{кол} + n_5 \cdot О_{дз};$$

[Оставьте те формы текущего контроля, которые предусмотрены в ОУП. сумма удельных весов должна быть равна единице: $\sum n_i = 1$] Способ округления накопленной оценки текущего контроля: [указывается способ – арифметический, в пользу студента, другое].

Результирующая оценка за дисциплину рассчитывается следующим образом:

1. Если дисциплина преподается один модуль:

$$О_{результ} = k_1 * О_{накопл} + k2 *\cdot О_{экз}$$

Способ округления накопленной оценки промежуточного (завершающего) контроля: [указывается способ – арифметический, в пользу студента, другое].

2. Если дисциплина преподается несколько модулей (например, 3):

$$О_{промежуточная\ i} = m_1 \cdot О_{текущая\ i\ этапа} + m_2 \cdot О_{промежуточный\ экзамен}$$

Где $О_{текущая\ i\ этапа}$ рассчитывается по приведенной выше формуле

$$О_{накопленная\ завершающая} = (О_{промежуточная\ 1} + О_{промежуточная\ 2} + О_{накопленная\ 3}) : на\ число\ модулей$$

Где $О_{промежуточная\ 1} + О_{промежуточная\ 2}$ – промежуточные оценки этапов 1 и 2,
а $О_{накопленная\ 3}$ – накопленная оценка последнего этапа перед завершающим экзаменом

Способ округления накопленной оценки промежуточного (завершающего) контроля в форме экзамена: [указывается способ – арифметический, в пользу студента, другое].
[Сумма удельных весов должна быть равна единице: $\sum m_i = 1$, при этом, $0{,}2 \le m_1 \le 0{,}8$ (согласно Положению об организации промежуточной аттестации и текущего контроля успеваемости студентов НИУ ВШЭ, утвержденному приказом НИУ ВШЭ от 19.08.2014 №6.18.1-01/1908-02**)**

ОПЦИОНАЛЬНО: На экзамене студент может получить дополнительный вопрос (дополнительную практическую задачу, решить к пересдаче домашнее задание), ответ на который оценивается в 1 балл.

[Оставьте те оценки, которые учитываются при выставлении результирующей оценки за промежуточный или завершающий контроль. Сумма удельных весов должна быть равна единице: $\sum k_i = 1$, при этом, **$0{,}2 \le k_1 \le 0{,}8$ После всех формул в обязательном порядке приводится способ округления полученного результата.**]

[Только для многомодульных дисциплин, по которым предусмотрен промежуточный контроль, укажите один из предложенных вариантов формирования оценки, которая идет в диплом]