

ПРИКЛАДНАЯ И КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА



- Компьютерная морфология
- Компьютерный синтаксис
- Компьютерное представление значений
- Распознавание и синтез речи
- Машинное обучение в лингвистике
- Корпусная лингвистика
- Машинный перевод
- Информационный поиск
- Извлечение информации
- Диалоги и чат-боты
- Анализ тональности
- Компьютерная текстология
- Квантитативная лингвистика:
что можно сосчитать в языке и речи?
- Речевое воздействие и манипулирование



URSS

ПРИКЛАДНАЯ И КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

Коллективная монография

Под редакцией
И. С. Николаева, О. В. Митрениной, Т. М. Ландо

Издание второе



URSS
МОСКВА

Прикладная и компьютерная лингвистика / Под ред. И. С. Николаева,
О. В. Митрениной, Т. М. Ландо. Изд. 2-е. — М.: ЛЕНАНД, 2017. — 320 с.

Вниманию читателей предлагается первое на русском языке практическое введение в современные лингвистические технологии. Из книги можно узнать о применении знаний о языке для решения прикладных задач. Монография позволяет найти ответы на базовые вопросы, возникающие у начинающего исследователя: как работают современные лингвистические технологии, где взять основные компоненты программ и что читать дальше для углубленного понимания.

Многие сложные научно-технические проблемы станут намного понятнее. Например, как заставить компьютер прочитать текст для слабовидящего человека. Или как сделать автоматический переводчик, чтобы договориться с торговцем на рынке в глухой провинции Китая. Или даже — как научить смартфон давать рекомендации, на какой фильм пригласить подругу.

Книга предназначена для преподавателей и студентов, для разработчиков программ по компьютерной обработке языка, для всех интересующихся многогранными возможностями современной прикладной лингвистики.

ООО «ЛЕНАНД».

117312, г. Москва, пр-т Шестидесятилетия Октября, д. 11А, стр. 11.

Формат 60×90/16. Печ. л. 20. Зак. № 3460.

Отпечатано в АО «Рыбинский Дом печати».

152901, г. Рыбинск, ул. Чкалова, д. 8.

ISBN 978–5–9710–4633–2

© ЛЕНАНД, 2016, 2017

22312 ID 228448



9 785971 046332



Все права защищены. Никакая часть настоящей книги не может быть воспроизведена или передана в какой бы то ни было форме и какими бы то ни было средствами, будь то электронные или механические, включая фотокопирование и запись на магнитный носитель, а также размещение в Интернете, если на то нет письменного разрешения владельца.

Содержание

Введение	12
Часть 1. Компоненты	14
Глава 1. Компьютерная морфология	14
1. Как найти слова	14
2. Каким может быть анализ слов	16
3. Лексическая неоднозначность	18
4. Анализ морфологии на основе правил	20
4.1. Что хранить в словарях?	20
4.2. Морфологические модули АОТ	22
4.3. Морфологический анализатор Rymorphy2 и словарь проекта OpenCorpora	24
4.4. Анализатор mystem	26
5. Статистические методы анализа слов	27
5.1. Статистическая частеречная разметка	27
5.2. Триграммная скрытая Марковская модель	29
5.3. Частеречная разметка незнакомых слов	32
<i>Литература</i>	<i>32</i>
<i>Электронные ресурсы</i>	<i>33</i>
Глава 2. Компьютерный синтаксис	35
1. Разные подходы к анализу синтаксических структур	35
1.1. Что такое парсинг	35
1.2. Грамматики зависимостей	36
1.3. Грамматики непосредственных составляющих	40
1.4. Комбинированные теории анализа предложения	43
2. Неоднозначность и проблема комбинаторного взрыва	44
3. Статистический парсинг	47
4. Современные синтаксические анализаторы: семь глаз и типы в цехе	48
4.1. Лингвистический процессор ЭТАП	48

4.2. DictaScope и AOT	50
4.3. Stanford NLP, RASP, OpenNLP.....	52
4.4. Link Grammar Parser	53
4.5. NLTK	55
5. Дальнейшие задачи	56
<i>Литература</i>	57
<i>Электронные ресурсы</i>	58
Глава 3. Компьютерное представление значений	59
1. О семантическом модуле	59
2. Модели представления знаний в компьютерной семантике	60
2.1. Виды семантических представлений.....	60
2.2. Сетевые модели	60
2.3. Концептуальные графы.....	62
2.4. Фреймы и сценарии.....	63
2.5. Современные разновидности семантических представлений.....	66
3. Формальные онтологии	67
3.1. Структура формальных онтологий	67
3.2. Классификация формальных онтологий.....	68
3.3. Особенности создания формальных онтологий.....	70
3.4. Языки представления и редакторы формальных онтологий.....	71
3.5. Методы автоматического построения формальных онтологий.....	72
3.6. Современные онтологические ресурсы	73
3.7. Применение формальных онтологий.....	75
3.8. Стандартизация и оценка качества формальных онтологий.....	77
4. Компьютерные тезаурусы.....	78
4.1. Какие бывают тезаурусы.....	78
4.2. Компьютерные тезаурусы типа WordNet.....	80
4.3. Компьютерные тезаурусы для русского языка	82
4.4. Надстройки к компьютерным тезаурусам	86
4.5. Прикладное использование компьютерных тезаурусов	87
5. Настоящее и будущее компьютерной семантики	88
<i>Литература</i>	89
<i>Электронные ресурсы</i>	92

Глава 4. Распознавание и синтез речи	94
1. Навстречу эпохе говорящих машин	94
2. Синтез речи.....	96
2.1. Методы синтеза	97
2.2. Устройство TTS-синтезатора речи	101
2.3. Модуль лингвистической обработки текста	102
3. Распознавание речи.....	105
3.1. Вариативность речи — главное препятствие для разработчиков систем распознавания речи.....	107
3.2. Основные типы современных систем распознавания речи.....	110
3.3. Лингвистический и статистический подходы к распознаванию речи	112
3.4. Скрытые Марковские модели	113
3.5. Как работает статистическая система распознавания речи?	114
4. Новые горизонты	117
<i>Литература</i>	118
<i>Электронные ресурсы</i>	119
Глава 5. Машинное обучение в лингвистике	121
1. Введение: Формализация задач машинного обучения.....	121
2. Методы машинного обучения	124
2.1. Метрические методы классификации	126
2.2. Статистические методы классификации	127
2.3. Линейные методы классификации	129
2.4. Регрессионные методы.....	130
2.5. Искусственные нейронные сети.....	131
2.6. Кластеризация.....	133
3. Заключение	135
<i>Литература</i>	136
<i>Электронные ресурсы</i>	137
Глава 6. Корпусная лингвистика	138
1. Корпусы вчера и сегодня	138
2. Основные свойства корпуса	139
2.1. Электронный.....	139
2.2. Репрезентативный	139

2.3. Размеченный	140
2.4. Прагматически ориентированный.....	141
3. Какие бывают корпуса	141
3.1. Параллельные корпуса	142
3.2. Корпусы устной речи	143
4. Разметка корпусов.....	143
4.1. Средства разметки	143
4.2. Лингвистическая разметка.....	144
5. Интернет как корпус	146
6. Сервис корпусного менеджера.....	148
7. Как сделать корпус самому	150
8. Корпусы как инструмент будущего	151
<i>Литература</i>	152
<i>Электронные ресурсы</i>	154

Часть 2. Направления 156

Глава 1. Машинный перевод..... 156

1. Три подхода к машинному переводу	156
2. Перевод на основе правил.....	158
2.1. Три способа перевода с помощью правил.....	158
2.2. Трансферный подход.....	159
2.3. Пример словарей и грамматик компании PROMT.....	160
3. Статистический машинный перевод	162
3.1. Главная формула перевода.....	162
3.2. Модель языка и цепи Маркова	163
3.3. Оценка максимального правдоподобия	166
3.4. Методы сглаживания	168
3.5. Модель перевода.....	170
4. Гибридный перевод.....	182
5. Методы оценки качества перевода	183
6. Некоторые современные системы машинного перевода.....	184
<i>Литература</i>	188
<i>Электронные ресурсы</i>	188

Глава 2. Информационный поиск	190
1. Где ищем?	190
2. Что ищем?	191
3. Как ищем?	192
3.1. Индекс	192
3.2. В идеальном мире	194
3.3. Тем временем в реальности	194
4. Что такое хорошо?	196
4.1. Релевантность, полнота, точность	196
4.2. Фильтрация и ранжирование	197
4.3. Факторы ранжирования	198
4.4. Оценки релевантности	199
4.5. Не все слова одинаково полезны	199
5. А где же лингвистика?	203
5.1. Стандартные запчасти	203
5.2. Расширения	204
5.3. Расстояния	206
5.4. Еще немного поисковой лингвистики	207
<i>Литература</i>	209
<i>Электронные ресурсы</i>	210
Глава 3. Извлечение информации	211
1. Какую информацию извлекаем?	211
2. Распознавание сущностей	214
2.1. Какие сущности извлекаем	214
2.2. Зависимость категории от контекста	215
2.3. Неоднозначность идентификации	216
2.4. Концептуальные сложности	216
2.5. Разрешение анафоры и кореферентности	217
2.6. Установление референта	218
2.7. Автоматические подходы	218
2.8. Использование экстратекстуальных сигналов	220
3. Извлечение отношений	222
3.1. Какие отношения извлекаем	222
3.2. Обучение моделей на размеченных текстах	223
3.3. Полуавтоматическое создание размеченного корпуса	224
3.4. Временное измерение	225

4. Извлечение событий	226
5. Для тех, кто хочет попробовать сам	230
<i>Литература</i>	231
<i>Электронные ресурсы</i>	232
Глава 4. Диалоги и чат-боты	233
1. Компьютер притворяется человеком	233
2. Особенности диалога на естественном языке	234
3. Архитектура диалоговых систем	235
3.1. Модуль понимания естественного языка	236
3.2. Диалоговый менеджер	236
3.3. Модуль генерации естественного языка	237
4. Как работают чат-боты	238
4.1. Имитация беседы	238
4.2. Язык AIML и другие подходы	239
5. Обучение диалоговых систем на реальных диалогах	241
6. Углубление диалога	242
<i>Литература</i>	243
<i>Электронные ресурсы</i>	244
Глава 5. Анализ тональности	245
1. Компьютер отслеживает чувства	245
2. С чего начинается оценка?	246
3. Как измерить тональность текста	248
3.1. Подход с использованием правил и словарей	248
3.2. Подход с использованием машинного обучения	253
4. Как это выглядит на практике	254
5. Оценка качества работы алгоритмов	255
<i>Литература</i>	256
<i>Электронные ресурсы</i>	258
Глава 6. Компьютерная текстология	259
1. Что такое текстология	259
2. Этапы текстологического исследования рукописной традиции	260

3. Компьютер в работе текстолога	265
3.1. Автоматическое сравнение рукописей.....	265
3.2. Компьютерная классификация рукописей.....	266
<i>Литература</i>	271
<i>Электронные ресурсы</i>	272
Глава 7. Квантитативная лингвистика: что можно сосчитать в языке и речи?	273
1. Буквы и звуки: как определить, на каком языке написан текст? — Дешифровка	273
2. Морфемы: как оценить сложность языка? — Типология.....	275
3. Части речи: можно ли определить, о чем текст? — Стилеметрия	277
4. Сто слов: как определить возраст языков? — Глоттохронология	279
5. Слова, слова, слова: сколько слов мы знаем и сколько нам нужно знать? — Частотные словари	283
6. Порядок, строй, парадигма: насколько стройна грамматика? — Квантитативная морфология	289
<i>Литература</i>	291
<i>Электронные ресурсы</i>	292
Глава 8. Речевое воздействие и манипулирование.....	294
1. Что такое речевое воздействие?.....	294
2. Кто и зачем изучает речевое воздействие?	296
3. Разновидности речевого воздействия.....	297
4. Языковое манипулирование: приемы и ресурсы.....	301
<i>Литература</i>	307
<i>Электронные ресурсы</i>	308
Указатель терминов.....	309