



Н. Н. ЛЕОНТЬЕВА

АВТОМАТИЧЕСКОЕ ПОНИМАНИЕ ТЕКСТОВ СИСТЕМЫ, МОДЕЛИ, РЕСУРСЫ

*Для студентов
лингвистических факультетов вузов*

1561816



Москва

ACADEMIA
2006

УДК 800(075.8)

ББК 81.1я73

Л478

Рецензенты:

доктор филологических наук, профессор, зав. кафедрой лингвистической семантики Московского государственного лингвистического университета

Б. Ю. Городецкий;

доктор филологических наук, профессор, главный научный сотрудник Института русского языка им. В.В.Виноградова РАН *В. М. Андриященко*

Леонтьева Н. Н.

Л478 Автоматическое понимание текстов: системы, модели, ресурсы: учеб. пособие для студ. лингв. фак. вузов / Нина Николаевна Леонтьева. — М.: Издательский центр «Академия», 2006. — 304 с.

ISBN 5-7695-1842-1

Учебное пособие обобщает опыт создания отечественных и зарубежных систем, реализующих автоматическое понимание текстов. Эти сложные «интеллектуальные» системы выделяются из множества систем, в которых просто используется автоматическая обработка текста, поскольку автора интересует именно качественный аспект понимания. Рассмотрены те компоненты процесса АПТ, которые могут быть заданы в вербальном виде. В основе пособия — идея «мягкого» понимания текста; представлена экспериментальная лингвистическая система ПОЛИТЕКСТ, осуществляющая гибкое соединение лингвистических и предметных знаний.

Для студентов лингвистических факультетов вузов. Может быть рекомендовано для тех, кто интересуется искусственным интеллектом, структурной и прикладной лингвистикой, информатикой.

УДК 800(075.8)

ББК 81.1я73

Оригинал-макет данного издания является собственностью Издательского центра «Академия», и его воспроизведение любым способом без согласия правообладателя запрещено

ISBN 5-7695-1842-1

© Леонтьева Н. Н., 2006

© Издательский центр «Академия», 2006

ОГЛАВЛЕНИЕ

| | |
|--|-----|
| Предисловие | 3 |
| Введение | 10 |
| Автоматическая обработка или понимание текста? | 10 |
| В центре внимания — лингвистический аспект | 10 |
| О модели | 11 |
| О проекте ПОЛИТЕКСТ | 12 |
| ПОЛИТЕКСТ — это система? | 12 |
| Место семантики | 14 |
| Глава 1. Взгляд «сверху» на системы автоматического понимания текста | 15 |
| § 1. Прикладная и теоретическая лингвистика | 15 |
| § 2. Что значит «автоматическое понимание текста» | 17 |
| § 3. Основные задачи и классы систем АПТ | 19 |
| § 4. Типы текстовых структур в системах АПТ | 21 |
| § 5. Состав компонентов стандартных систем АПТ | 27 |
| § 6. Модель «мягкого понимания» текста | 29 |
| § 7. Синтез информационного и лингвистического подходов | 30 |
| § 8. Процесс понимания как взаимодействие текстов | 32 |
| Глава 2. Машинный перевод как среда создания систем автоматического понимания текста | 36 |
| § 9. Об истории СМП | 36 |
| § 10. Периодизация и классификация СМП | 38 |
| § 11. Лингвистическое обеспечение СМП | 40 |
| § 12. Внешняя и внутренняя оценка СМП | 42 |
| § 13. Нерешенные проблемы автоматического понимания и перевода | 43 |
| § 14. Новая парадигма СМП | 44 |
| § 15. Включение предметной области как задача информационно-переводческой системы | 45 |
| Глава 3. Компонент первичного анализа текста | 49 |
| § 16. Состав компонента первичного анализа текста | 51 |
| Препроцессор: подготовка массива | 52 |
| Препроцессор: создание внешней дескрипции документа | 52 |
| Стандарты оформления документов | 55 |
| § 17. Собственно графематический анализ | 56 |
| § 18. Макросинтаксический анализ | 58 |
| § 19. Проблема анализа прерванных высказываний | 60 |
| Глава 4. Компонент морфологического анализа | 64 |
| § 20. Подходы к МорфАн | 65 |
| МорфАн со словарем основ и словарем окончаний | 66 |
| МорфАн только со словарем окончаний | 68 |
| МорфАн «по аналогии» | 69 |
| МорфАн со словарем словоформ в системе ПОЛИТЕКСТ | 70 |
| § 21. Семантические проблемы в МорфАн | 74 |
| Глава 5. Синтаксический компонент | 78 |
| § 22. Проблема синтаксической омонимии при анализе | 79 |
| § 23. Модели автоматического СинАн | 81 |
| § 24. Составляющие синтаксического компонента | 83 |
| § 25. О некоторых отечественных реализациях СинАн | 83 |
| § 26. Синтаксические процессоры в ИЛМ | 86 |
| Синтаксический компонент системы ФРАП | 87 |
| Синтаксический компонент системы ПОЛИТЕКСТ | 92 |
| Синтаксис в системе ДИАЛИНГ | 95 |
| § 27. Трудности, связанные с развитием синтаксического компонента | 96 |
| Глава 6. Локальный семантический анализ текста | 101 |
| § 28. Три структурных отображения текста: семантическое, информационное, когнитивное | 102 |
| § 29. Состав семантического компонента | 104 |
| § 30. Метаязык семантических структур | 105 |
| Функции и структура ИЯП | 105 |
| Смысловая грамматика | 107 |
| § 31. О единицах СемАн | 109 |
| § 32. Этапы локального СемАн текста | 112 |
| «Прямая» семантическая интерпретация СинП | 113 |
| Семантическая интерпретация сильных связей | 114 |
| Семантическая интерпретация слабых связей | 119 |
| Проблема неполных актантажных структур | 123 |
| Глава 7. Глобальный семантический анализ и сжатие текста | 128 |
| § 33. Связность и смысловое сжатие текста | 128 |
| § 34. Информационный синтез значимых для текста единиц | 133 |
| § 35. Ситуация и ситуативное представление | 134 |
| § 36. Грамматика текстовых ситуаций | 139 |
| § 37. Критерии полноценности узлов и связей СемП | 141 |
| § 38. О полезных свойствах текста и его структур, на которые опираются механизмы глобального анализа | 141 |
| § 39. Гипертекст как информационное пространство текстов | 142 |
| Глава 8. Учет специальных знаний в системах автоматического понимания текста | 146 |
| § 40. Проблема предметной области | 146 |

| | |
|--|------------|
| § 41. Способы вовлечения специальных знаний в системы автоматического понимания естественного текста | 148 |
| § 42. Тезаурусы | 149 |
| WordNet, EuroWordNet | 150 |
| Некоторые отечественные тезаурусы | 152 |
| RuТез | 153 |
| Синонимические ряды дескрипторов RuТез | 155 |
| Многозначные термины в RuТез | 157 |
| Система отношений между дескрипторами RuТез | 158 |
| § 43. Другие ПО-ориентированные словари и системы | 160 |
| Словарь-тезаурус энциклопедических функций | 160 |
| Описание ситуаций и схем ПО для одной фактографической ИПС | 164 |
| Аппарат семантических признаков в отраслевом словаре | 168 |
| Глава 9. Information Extraction и другие информационные модели | 174 |
| § 44. Автоматическое индексирование текстов | 175 |
| § 45. Автоматическое реферирование/фрагментирование текстов .. | 178 |
| § 46. Системы «вопрос-ответ» | 179 |
| § 47. Тематический анализ потока текстов | 180 |
| Создание тематического представления текста по тезаурусу | 181 |
| Разрешение неоднозначности терминов RuТез | 182 |
| Построение аннотации | 184 |
| § 48. Системы автоматического извлечения знаний из текстов | 185 |
| Глава 10. Системы генерации текста | 193 |
| § 49. Компоненты СГТ | 194 |
| § 50. Схемы процесса генерации текстов | 196 |
| § 51. Использование риторических структур в СГТ | 198 |
| § 52. Описание системы FoG | 201 |
| § 53. Сравнение систем МП и ГТ | 203 |
| § 54. Концептуальные vs. семантические структуры текста в СГТ | 204 |
| Глава 11. Концепция Базы текстовых фактов | 208 |
| § 55. Этапы построения и единицы БТФ | 209 |
| § 56. О важности создания БТФ для общественных наук | 213 |
| § 57. Роль и функции спецтранслятора в модели АПТ | 216 |
| § 58. Адаптация системы АПТ к новым ПО | 218 |
| § 59. Об универсальности лингвистического транслятора | 221 |
| § 60. Об информационной относительности в системе АПТ | 223 |
| § 61. Схема построения многоязыковой БТФ | 224 |
| Глава 12. Семантические словари: структура и состав информации | 227 |
| § 62. Словарный комплекс РУСЛАН | 227 |
| § 63. Типы входов в словарь | 228 |

| | |
|---|-----|
| § 64. О метаязыке словарных описаний | 229 |
| § 65. Категоризация лексики | 229 |
| § 66. Состав информации в семантическом словаре лексем | 232 |
| § 67. Подробное описание полей словаря лексем | 233 |
| Зона МОРФ (Морфологические данные) | 233 |
| Зона СИН (Синтаксические сведения) | 235 |
| Зона СЕМ (Семантические описания) | 235 |
| Зона ВАЛЕНТ (Семантические валентности) | 240 |
| Зона СИТ (Ситуации) | 244 |
| Зона ИНФ (Описание слова как единицы информационной структуры текста и тезауруса) | 247 |
| Зона ПРАГМ (прагматика) | 248 |
| Зона ЛЕКС (Лексическая сочетаемость) | 249 |
| Зона ЭКВ (Иноязычные эквиваленты). Поля АНГ, ФР, БОЛГ | 251 |
| Зона КОММ (Комментарии составителя) | 251 |
| § 68. Состав информации в словаре отношений | 251 |

| | |
|--|------------|
| Глава 13. Корпусная лингвистика и другие лингвистические ресурсы для систем АПТ | 258 |
| § 69. АРМ лингвиста, переводчика, редактора | 258 |
| § 70. Корпусная лингвистика | 259 |
| § 71. Аннотированный корпус | 261 |
| § 72. Методы анализа в КЛ | 263 |
| § 73. Теоретические позиции КЛ | 265 |
| § 74. КЛ, системы АПТ, лингвистика | 266 |
| § 75. Многоязыковая корпусная лингвистика | 267 |
| Заключение | 273 |
| Вопросы и задания ко всем главам | 277 |
| Список наиболее употребительных сокращений | 281 |
| Приложение 1 | 282 |
| Приложение 2 | 284 |
| Приложение 3 | 287 |
| Приложение 4 | 288 |
| Приложение 5 | 289 |
| Приложение 6 | 290 |
| Приложение 7 | 291 |
| Приложение 8 | 292 |
| Приложение 9 | 293 |
| Приложение 10 | 294 |
| Приложение 11 | 295 |
| Приложение 12 | 296 |
| Приложение 13 | 297 |