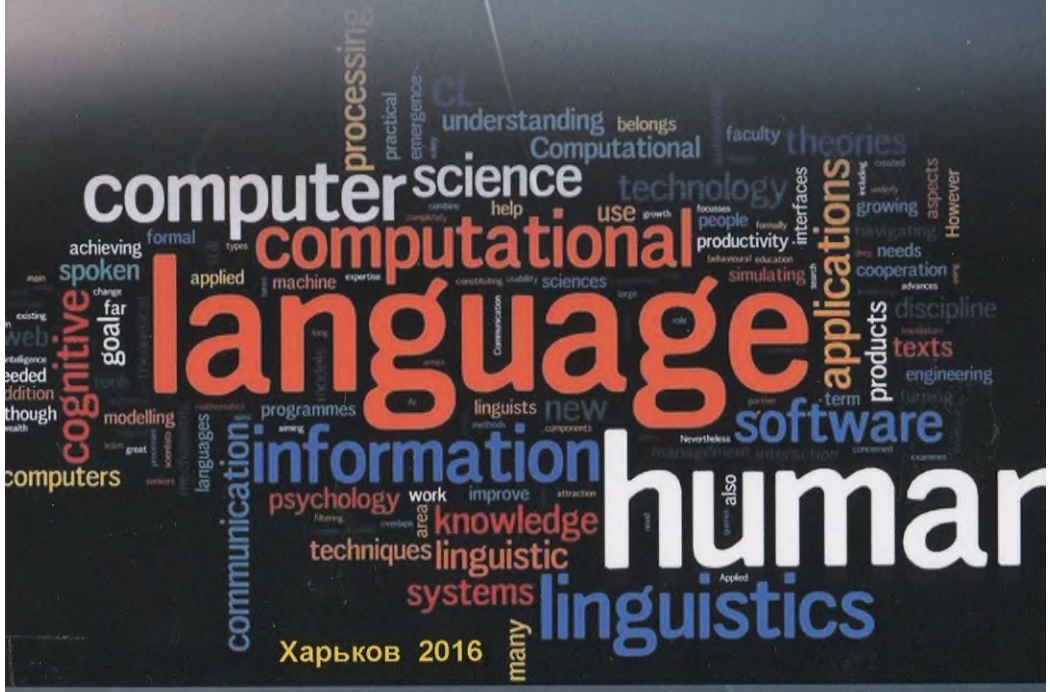
МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ УКРАИНЫ
НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ХАРЬКОВСКИЙ ПОЛИТЕХНИЧЕСКИЙ ИНСТИТУТ»

Н.Ф. ХАЙРОВА, Н.В. ШАРОНОВА

ИНФОРМАЦИОННО-ЛИНГВИСТИЧЕСКИЕ ТЕХНОЛОГИИ ЭКСТРАКЦИИ И ИДЕНТИФИКАЦИИ ГЛУБИННЫХ ЗНАНИЙ В ТЕКСТАХ



Харьков 2016

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ УКРАИНЫ
НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ХАРЬКОВСКИЙ ПОЛИТЕХНИЧЕСКИЙ ИНСТИТУТ»

Н. Ф. ХАЙРОВА, Н. В. ШАРОНОВА

ИНФОРМАЦИОННО-ЛИНГВИСТИЧЕСКИЕ ТЕХНОЛОГИИ ЭКСТРАКЦИИ И ИДЕНТИФИКАЦИИ ГЛУБИННЫХ ЗНАНИЙ В ТЕКСТАХ

Харьков
НТУ «ХПИ»
2016

УДК 004.934
ББК 32.811
Х 12

Рецензенты:
М. Д. Годлевский, д-р техн. наук, проф. НТУ «ХПИ» (г. Харьков);
Д. В. Ландэ, д-р техн. наук, ст. науч. сотр Института проблем регистрации информации (ИПРИ) НАН Украины (г. Киев)

Публикуется по решению ученого совета университета,
протокол № 11 от 22 декабря 2014 г.

Хайрова Н. Ф.
Х12 Информационно-лингвистические технологии экстракции и идентификации глубинных знаний в текстах: монография / Н. Ф. Хайрова, Н. В. Шаронова. - Х.: ФЛП Коряк С. Ф., 2016. - 205 с. - Рус. яз.

ISBN 978-966-97519-6-6

В монографии проблемы экстракции и идентификации глубинных знаний освещены в рамках задач, решаемых компьютерной лингвистикой. Рассмотрены вопросы моделирования семантических парадигматических отношений элементов естественно-языковой системы, многомерного представления пространства знаний текстового репозитория, описаны технологии семантически ориентированной идентификации тональности разноязычных текстов и извлечения фактографической информации из слабоструктурированных текстовых источников.

Для студентов, аспирантов, специалистов в области интеллектуальных компьютерных систем, прикладной компьютерной лингвистики и информационных технологий.

Ил. 44. Табл. 16. Библиогр.:279 наим.

УДК 004.934
ББК 32.812

ISBN 978-966-97519-6-6

© Хайрова Н. Ф., 2016
© Шаронова Н. В., 2016
© ФЛП Коряк С.Ф., 2016

СОДЕРЖАНИЕ

Список используемых сокращений..... 6
Введение..... 7

Раздел 1. СИСТЕМНЫЙ АНАЛИЗ ПРОБЛЕМЫ ИЗВЛЕЧЕНИЯ ЗНАНИЙ ИЗ СЛАБОСТРУКТУРИРОВАННОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ

1.1. Современные подходы и системы автоматизированной обработки текстов..... 9
1.2. Современное состояние исследований в области идентификации и представления знаний..... 13
1.3. Состояние и перспективы развития приложений лингвистического процессора..... 18
1.4. Существующие модели и методы извлечения фактографических знаний..... 23
1.5. Проблемы идентификации знаний в слабоструктурированных текстовых информационных потоках..... 27

Раздел 2. КОНЦЕПЦИЯ ИДЕНТИФИКАЦИИ СМЫСЛА ЭЛЕМЕНТОВ СВЯЗНОГО ТЕКСТА

2.1. Определение области исследования семантических отношений сложной языковой системы как междисциплинарной области системно-кибернетических знаний..... 34
2.2. Решение проблемы экстракции и идентификации знаний в рамках задач, решаемых компьютерной лингвистикой..... 37
2.3. Формальная экстракция и локализация глубинных и мягких знаний связного текста..... 40
2.4. Представление естественного языка как сложной слабоформализуемой системы..... 44
2.5. Концептуальная схема идентификации смысла элементов сложной иерархической языковой системы..... 48

Раздел 3. РАЗРАБОТКА ФОРМАЛЬНОЙ МОДЕЛИ ИДЕНТИФИКАЦИИ СЕМАНТИЧЕСКИХ ОТНОШЕНИЙ ЭЛЕМЕНТОВ ЕСТЕСТВЕННО-ЯЗЫКОВОЙ СИСТЕМЫ

3.1. Базовые логико-алгебраические средства теории интеллекта..... 54
3.2. Формальная структура и метод построения бинарной логической сети..... 57
3.3. Основные категории модели семантической классификации знаковых смысловых единиц..... 61

3.4. Математическая модель корреляции концепта знака лингвистической единицы и инсайтного понимания элемента связного текста, содержащего данный знак.....64
3.5. Введение понятия репрезентативного элемента и репрезентативного отношения семантического поля лингвистических смысловых единиц 70

Раздел 4. МЕТОД ФОРМАЛИЗАЦИИ СЕМАНТИЧЕСКИХ ПАРАДИГМАТИЧЕСКИХ ОТНОШЕНИЙ ЛИНГВИСТИЧЕСКИХ СМЫСЛОВЫХ ЕДИНИЦ В НЕСТРУКТУРИРОВАННОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ

4.1. Обоснование выбора значимых для представления знаний корреляций семантических полей 74
4.2. Логико-лингвистическая модель извлечения фактов из слабоструктурированной текстовой информации.....79
4.3. Модель экстракции и идентификации глубинных знаний из потоков текстовой информации..... 84
4.4. Принципы работы логической сети, формализующей идентификацию знаний из текстов информационной системы.....88
4.5. Многомерное представление пространства знаний текстового репозитория..... 94

Раздел 5. ЛОГИКО-ЛИНГВИСТИЧЕСКИЕ МОДЕЛИ ОПИСАНИЯ СЕМАНТИЧЕСКИХ ОТНОШЕНИЙ КОНЦЕПТОВ ФРАЗЫ И СВЕРХФРАЗОВОГО ЕДИНСТВА

5.1. Классификация ошибок машинного перевода, основанная на процедурах лингвистического процессора.....100
5.2. Логико-лингвистическая модель семантических отношений концептов сверхфразового единства.....105
5.3. Использование логической сети для семантического анализа связных фрагментов текста.....110
5.4. Логико-лингвистическая модель формального определения семантических падежей партиципантов предложения..... 115
5.5. Использование математического аппарата предикатных категорий для моделирования семантики сверхфразовых единств..... 121

Раздел 6. ИНФОРМАЦИОННО-ЛИНГВИСТИЧЕСКАЯ ТЕХНОЛОГИЯ СЕМАНТИЧЕСКИ ОРИЕНТИРОВАННОЙ ИДЕНТИФИКАЦИИ ТОНАЛЬНОСТИ РАЗНОЯЗЫЧНЫХ ТЕКСТОВ

6.1. Этапы информационно-лингвистической технологии экстракции новых знаний из слабоструктурированной текстовой информации 127

6.2. Особенности экстракции и идентификации
знаний из Web-контента.....133
6.3. Поэтапное представление информационно-лингвистической
технологии решения задачи Opinion Mining.....137
6.4. Структура базового многоязыкового тезауруса идентификации
тональности, учитывающего смысловые эквиваленты.....144
6.5. Выявление семантических эквивалентов
в слабоструктурированных текстах.....147

Раздел 7. ВНЕДРЕНИЕ РЕЗУЛЬТАТОВ ИССЛЕДОВАНИЯ В ПРАКТИКУ СОЗДАНИЯ СИСТЕМ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

7.1. Практическая реализация полученных результатов
при построении таксономии предметизатора
полнотекстовых электронных статей библиотеки.....154
7.2. Практическая реализация разработанных технологий
в системах электронного документооборота.....158
7.3. Использование логико-алгебраической модели падежной
грамматики для снятия семантической омонимии
в системах украинско-английского машинного перевода.....162
7.4. Практическое использование модели извлечения
фактографической информации из слабоструктурированных
текстов в системе формирования библиографических описаний
полнотекстовых документов научной библиотеки.....165
7.5. Формальная модель оценивания качества экстракции и
идентификации знаний из слабоструктурированной
текстовой информации.....170
7.6. Экспериментальная оценка эффективности и качества
разработанных лингвистических технологий идентификации
знаний в слабоструктурированной текстовой информации.....175

Заключение.....179
Список использованных источников.....181