Тезаурусы в задачах информационного поиска
Лукашевич Н.В.
М.: Изд-во МГУ, 2011. – 512 с., ил.

Область современного информационного поиска чрезвычайно разнообразна. Она включает такие задачи, как собственно поиск информации, фильтрация, рубрикация и кластеризация документов, поиск ответов на вопросы, автоматическое аннотирование документа и группы документов, поиск похожих документов и дубликатов, сегментирование документов и многое другое. Когда подобные операции выполняет человек, ему необходимо выявить основное содержание документа, его основную тему и подтемы, и для этого обычно используется большой объем знаний о языке, мире, организации связного текста.

Подавляющее число современных методов обработки неструктурированной информации решают эти задачи на основе минимальных предварительных знаний и базируются на моделях текста как набора слов (“bag of words”), предлагая изощренные методы учета частотностей встречаемости слов в предложении, тексте, наборе документов, совместной встречаемости слов и т.п. Пословные модели не учитывают такие языковые явления, как синонимия, многозначность, существование лексических отношений между словами.

Книга Лукашевич Н.В. «Тезаурусы в задачах информационного поиска» посвящена описанию опыта автора по созданию сверхбольших лингвистических ресурсов для автоматической обработки текстов в рамках современных информационных технологий и сопоставлению созданных ресурсов и технологий с подобными проектами, развиваемыми в мире.

Книга делится на два раздела.

В первом разделе (части 1-3) описываются различные подходы к созданию больших лингвистических ресурсов на примере конкретных проектов, а также подробно рассматриваются различные алгоритмы и системы, которые используют эти ресурсы для решения различных задач информационного поиска. При описании алгоритмов особое внимание уделяется методам оценки их качества, достигнутым показателям, которые указывают на то, удалось или нет разработчикам ресурсов и алгоритмов достигнуть лучшего качества по сравнению с пословными статистическими методами.

Во второй разделе книги (части 4-6) описываются принципы разработки лингвистического ресурса русского языка тезауруса РуТез и эксперименты по применению этого тезауруса в различных задачах обработки текстов для приложений информационного поиска. Описывая собственные алгоритмы, автор также уделяет большое внимание экспериментам, которые показывают, насколько качественно удается решать конкретные задачи на базе тезаурусных знаний.

В каждом из двух разделов книги выделяются части, которые подразделяются на главы. Первая часть первого раздела книги посвящена описанию различных видов тезаурусов, включая тезаурус Роже, информационно-поисковые тезаурусы, тезаурусы типа WordNet.

Во второй части книги рассматриваются основные положения современных онтологических исследований, принципы создания онтологических ресурсов. Особое внимание уделяется принципам установления онтологических отношений, которые нужны для создания ресурсов в различных предметных областях. В их число входят отношения *класс-подкласс*, *часть-целое*, отношения онтологической зависимости.

Следующая, третья часть, описывает применение тезаурусов и онтологий в конкретных приложениях информационного поиска. Здесь

рассматриваются такие системы, как собственно информационный поиск, системы автоматической рубрикации, вопросно-ответные системы, алгоритмы разрешения лексической многозначности, алгоритмы установления лексической связности в тексте, алгоритмы автоматического аннотирования текстов.

Каждая глава этой части строится схожим образом. Сначала описывается общая постановка задачи, некоторые теоретические положения и (или) основные статистические пословные алгоритмы, а также меры измерения качества решения задачи, а далее излагаются методы и результаты применения тезаурусов и онтологий в данной задаче.

С четвертой части начинается второй раздел книги, посвященный рассмотрению собственных ресурсов, создаваемых под руководством и при непосредственном участии автора, и экспериментов с ними. В этой части рассматриваются основные принципы построения тезауруса РуТез, методы описания понятий, языковых выражений, тезаурусных отношений, способы отражения разных значений слов, терминов, языковых выражений, описание синонимичности языковых выражений.

В пятой части книги рассматриваются эксперименты и приложения, основанные на знаниях, описанных в тезаурусе РуТез. В число этих приложений входят: информационный поиск, автоматическая рубрикация текстов, автоматическое аннотирование отдельного текста и совокупности сходных текстов, автоматической разрешение лексической многозначности, построение лексических цепочек и тематического представления связного текста.

Последняя, шестая часть книги посвящена основным направлениям развития тезауруса РуТез, а также технологии разработки других ресурсов, которые были созданы на основе тезауруса РуТез, а именно, принципы устройства и современное состояние Онтологии по естественным наукам и технологиям (ОЕНТ).

Книга предназначена для специалистов, научных работников, аспирантов и студентов, интересующихся вопросами автоматической обработки текстов, применения в информационном поиске лингвистических ресурсов, а также информационным поиском в целом, практическими вопросами применения онтологий.

Для читателей, не знакомых с теориями, применяемыми в компьютерной лингвистике, семантике, с одной стороны, или с теорией и практикой информационного поиска, тестирования информационно-поисковых систем, с другой стороны, в книге представлен необходимый для понимания материал, насколько это было возможно в рамках одной книги.