Системы обработки и хранения корпусов: Manatee, Bonito и Word Sketches для чешского языка

Павел Рыхли, Павел Смрж

Университет им. Масарика, Факультет информатики

Ботаницка 68a, 60200 Брно, Чехия

{pary,smrz@fi.muni.cz}

Аннотация

Данная статья посвящена описанию системы обработки корпусов Manatee, предназначенной для работы с большими и супер-большими (объемом более 1 миллиарда словоупотреблений) корпусами текстов, главными особенностями которой являются быстрая обработка разнообразных сложных запросов к корпусу, а также вычисление различных статистических параметров корпуса, запросов и пр. В статье обсуждаются основные функции и возможности Manatee, рассматривается один из графических интерфейсов системы - Bonito. Сложная статистическая обработка текстовых данных с помощью Manatee представлена на примере вычисления так называемых «словесных описаний» (Word Sketches). В статье особое внимание уделяется построению Word Sketches для чешского языка, а также решению проблем, связанных со свободным порядком слов. В заключении проводится сравнение процедур построения Word Sketches и традиционных методов автоматического выявления семантических отношений.