Лингвистическая онтология тезаурус РуТез и приложения автоматической обработки текстов

Лукашевич Наталья Валентиновна

Ведущий научный сотрудник НИВЦ МГУ

louk_nat@mail.ru

Доклад посвящен описанию структуры тезауруса русского языка РуТез, который представляет собой крупнейшую лингвистическую онтологию, используемую для автоматической обработки текстов. Структура тезауруса основана на традициях построения трех типов компьютерных ресурсов:

- традиционных информационно-поисковых тезаурусов,
- тезаурусов типа WordNet,
- формальных онтологий.

При автоматической обработке текстов на основе тезауруса РуТез строится так называемое тематическое представление, в котором основное содержание текстов моделируется совокупностями близких по смыслу понятий, упомянутых в тексте (тематических узлов), и отношений между ними.

Построенное тематическое представление используется для порождения концептуального индекса текстов, который может загружаться в информационно-поисковые системы и служить для концептуального индексирования (например, в информационно-поисковой системе УИС РОССИЯ www.uisrussia.ru), которое дает возможность поиска на основе тезауруса, автоматического расширения запросов, визуализации результатов поиска

Также тезаурус РуТез и построенное тематическое представление текстов используется в таких приложениях, как автоматическая рубрикация (классификация) текстов, автоматическое аннотирование одного и многих документов, автоматическая кластеризация документов. В докладе будут рассмотрены результаты тестирования подходов, основанных на тезаурусе РуТеЗ, в том числе и в рамках таких семинаров, как SUMMAC (Summarization Conference) и РОМИП (Российский семинар по методам информационного поиска).