Методы и средства построения программных систем для анализа текста с использованием лингвистических шаблонов

Большакова Елена Игоревна, Ефремова Наталья Эрнестовна, Носков Алексей Анатольевич
Кафедра алгоритмических языков
bolsh@cs.msu.su, nvasil@list.ru, alexey.noskov@gmail.com

Современные прикладные системы по автоматической обработке текстов на естественном языке, решающие задачи извлечения информации из текстов, реферирования и аннотирования текстов и др. [1], базируются в основном на частичном синтаксическом анализе для распознавания в тексте определенных языковых конструкций. Стремительный рост количества таких приложений делает актуальным разработку специализированных средств, упрощающих их построение; к числу таких средств относятся языки формализации лингвистических свойств распознаваемых конструкций и поддерживающие их инструментальные системы.

Для автоматической обработки русскоязычных текстов был предложен язык лексико-синтаксических шаблонов LSPL и разработаны поддерживающие его программные средства [2]. Шаблон специфицирует входящие в распознаваемую конструкцию слова с учетом их морфологических характеристик и условий грамматического согласования. Ядром разработанного программного комплекса является компонент поиска и выделения в тексте конструкций по их лексико-синтаксическим шаблонам.

В докладе характеризуются особенности построения приложений по обработке текстов с использованием языка LSPL и созданного программного комплекса, включающего также средства интеграции языка в приложения и визуальную среду для просмотра и анализа текстов с использованием шаблонов. Шаблоны подготавливаются лингвистом или специалистом по предметной области анализируемых текстов; набор шаблонов хранится отдельно от кода приложения и может модифицироваться без участия программиста. Программный комплекс был применен для построения нескольких приложений, в том числе процедур терминологического анализа научно-технического текста и модуля генерации программных тестов по комментариям программного кода [3].

В докладе обсуждается также дальнейшее развитие исследований в направлении построения инструментальной среды для быстрой разработки более широкого класса приложений, которая допускает использование сторонних модулей анализа текста на базе общей модели текстовых данных и позволяет настраивать процесс выполнения приложений.

Литература

1. Grishman R. Information extraction. In: The Oxford Handbook of Computational Linguistics. Mitkov R. (ed.). Oxford University Press, 2003, p. 545-559.
2. Большакова Е.И., Носков А.А. Система для поиска и выделения конструкций в текстах на естественном языке // Двенадцатая национ. конференция по искусств. интеллекту с междунар. участием (КИИ-2010): Труды конференции. Т. 4. — М., 2010, с.63-71.
3. Bolshakova E., Efremova N., Noskov A. LSPL-Patterns as a Tool for Information Extraction from Natural Language Texts // New Trends in Classification and Data Mining. K.Markov et al. (Eds.). Sofia, ITHEA, 2010, p. 110-118.