ПРОГРАММНЫЕ СРЕДСТВА АНАЛИЗА ТЕКСТОВ НА ОСНОВЕ ЛЕКСИКО-СИНТАКСИЧЕСКИХ ШАБЛОНОВ

МОРОЗОВ Сергей Сергеевич

E-mail: sergej_morozov@rambler.ru

Научный руководитель: БОЛЬШАКОВА Елена Игоревна

Кафедра Алгоритмических Языков

Данная дипломная работа посвящена разработке формального языка для декларативного описания лексических и синтаксических особенностей языковых конструкций в системах автоматической обработки текстов на естественном языке (ЕЯ).

При решении различных задач компьютерной лингвистики часто применяется поверхностно-синтаксичекий анализ, на основе которого происходит выделение языковых конструкций (например, именных словосочетаний) по их описанию, сделанному формальными средствами. Проведенное в данной работе исследование формальных средств описания и механизмов поиска в системах автоматической обработки текстов на русском языке показало их основные недостатки, главным из которых является невозможность явного описания грамматической связи согласования, типичной для выделяемых русских словосочетаний.

Разработанный в данной дипломной работе язык LSPL (LexicoSyntactic Pattern Language) основан на понятии лексико-синтаксического шаблона [1]. Под шаблоном понимается структурный образец, отображающий лексические и поверхностно-синтаксические свойства описываемых фрагментов текста.

Лексико-синтаксический шаблон может включать различные элементы: элементы-строки, элементы-слова, конструкции повторения и экземпляры шаблонов. Элемент-строка позволяет в точности задать некоторую символьную строку. Элемент-слово позволяет описать некоторое слово текста, указав соответствующую лексему, символьный образец, морфологические характеристики слова. Конструкция повторения задает повторяющие однотипные фрагменты текста. Экземпляр шаблона позволяет использовать ранее описанные шаблоны и тем самым постепенно описывать все более сложные конструкции естественного языка.

Важной составляющей шаблона являются условия согласования, записываемые после описания всех элементов и позволяющие задать согласование элементов шаблона как по всем их общим морфологическим характеристикам, так и по некоторым из этих характеристик.

Для автоматического поиска в текстах на ЕЯ фрагментов, описанных в виде лексико-синтаксического шаблона, разработан алгоритм поиска и реализованы основанные на нем программные средства. Основной составляющей реализованных программных средств является LSPL-обработчик, который позволяет создавать и модифицировать набор шаблонов и выполняет поиск всех описываемых некоторым шаблоном фрагментов текста,. Особенностью разработанного алгоритма является тесное взаимодействия двух процессов: поиска некоторого фрагмента текста и сопоставления шаблона с этим текстовым фрагментом.

Разработанный язык LSPL покрывает все основные возможности по описанию лексико-синтаксических свойств, реализованные в системах автоматической обработки русских текстов, а также позволяет в явном виде описывать связь грамматического согласования. Язык пригоден равно

- как язык записи запросов на поиск исследуемых конструкций в текстах, формулируемых на основе их словарного состава и несложных грамматических условий;
- как способ формальной записи специфических языковых конструкций для их представления в системе автоматической обработки текстов различных стилей.

Запись на языке LSPL является декларативной и обеспечивает достаточно естественный для пользователя-лингвиста способ описания выделяемых фрагментов текста

Литература:

1. Большакова Е.И., Васильева Н.Э., Морозов С.С. Лексико-синтаксические шаблоны для автоматического анализа научно-технических текстов // Десятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2006. Труды конференции в 3-х томах. Т. 2. – М.: Физматлит, 2006, с.506-524.