Программная поддержка языка лексико-синтаксических шаблонов LSPL

Носков Алексей Анатольевич
студент кафедры алгоритмических языков
e–mail: alexey.noskov@gmail.com
Научный руководитель – к.ф.-м.н., доцент Большакова Елена Игоревна

Дипломная работа посвящена проблемам автоматического выявления различных конструкций в текстах на естественном языке.

В настоящее время в области создания систем автоматической обработки естественного языка актуальна задача выделения в текстах определенных языковых конструкций, например, согласованных именных словосочетаний (*усталое осеннее солнце, уходящий поезд*), глагольных групп (*шел по тротуару, писать стихи*), а также более сложных конструкций, характерных например для текстов научно-технического стиля (*под А будем понимать В, предположим, что С*) и т.п. До сих пор задача такого выделения обычно решалась каждый раз заново в условиях конкретного приложения по автоматической обработке текста и для отдельных типов языковых конструкций. В данной работе предлагается новый метод, позволяющий осуществлять выделение в тексте достаточно широкого круга языковых конструкций, описанных в виде шаблонов языка LSPL [1], расширенного средствами задания словарной информации.

Язык LSPL, используемый в качестве гибкого средства описания конструкций естественного языка для их автоматического выделения в тексте с учетом особенностей русского языка, позволяет записывать конструкции в виде так называемых лексико-синтаксических шаблонов. Шаблон языка LSPL в общем случае состоит из нескольких альтернатив, описывающих различные варианты выделяемой конструкции и состоящих из последовательности элементов, представляющих слова с их морфологическими характеристиками и условиями грамматического согласования. Язык включает средства записи повторяющихся элементов описываемой конструкции и позволяет задавать фрагменты конструкции с помошью уже определенных шаблонов. Например, шаблон

```
NG = {A} N1 <A=N1> [NG2<c=gen>]
```

описывает именную группу, состоящую из последовательности прилагательных (`{A}`), существительного (`N1`), согласованного с этими прилагательными (`<A=N1>`), и опциональной именной группы в родительном падеже (`[NG2<c=gen>]`). Такому шаблону соответствует фразы вида «*белый кот*», «*долгая зимняя ночь*» и «*тоненькая струйка дыма далекого пожара*».

Разработанный метод базируется на представлении обрабатываемого текста в виде специального графа, ребра которого представляют синтаксические интерпретации фрагментов текста. Для одного и того же фрагмента текста в графе может содержаться несколько ребер, представляющих его различные интерпретации. Первоначально в графе хранятся синтаксические интерпретации слов текста, а затем добавляются промежуточные результаты анализа — интерпретации фрагментов текста. Для ускорения поиска в графе используются индексы ребер графа: индекс частей речи, индекс шаблонов и индекс слов текста. Для сокращения множества результатов поиска используется специальная группировка ребер графа.

На основе метода реализован комплекс программных средств, основными компонентами которого являются:

- Центральный компонент выделения конструкций по шаблонам;
- Программный интерфейс для использования возможностей комплекса внешними приложениями на языке Java;
- Консольные утилиты, позволяющие производить отладку шаблонов и автоматическое выделение в тексте конструкций по LSPL-шаблонам;
- Графический пользовательский интерфейс, позволяющий загружать текст из различных текстовых документов и производить автоматическое выделение языковых конструкций на базе LSPL-шаблонов.

Важными особенностями реализованного программного комплекса является его кроссплатформенность, а также возможность его простой интеграции с другими приложениями автоматической обработки текста, разработанными на языках C++ или Java.

Комплекс был успешно протестирован при написании и отладке шаблонов типичных конструкций определения новых терминов, употребляемых в русскоязычных текстах научно-технического стиля. Разработанные программные средства доступны для свободного использования (расположены на сайте lspl.ru, посвященном языку лексико-синтаксических шаблонов LSPL).

Литература

1. Большакова Е.И., Баева Н.В., Бордаченкова Е.А., Васильева Н.Э., Морозов С.С. *Лексико-синтаксические шаблоны в задачах автоматической обработки текстов* // Компьютерная лингвистика и интеллектуальные технологии: Труды Межд. конф. Диалог '2007 – М.: Изд. центр РГГУ, 2007, стр. 70-75.