Генерация лексико-синтаксических шаблонов на основе извлекаемых из текста конструкций

Шариков Георгий Феликсович
студент кафедры алгоритмических языков
e-mail: egos_@mail.ru
Научный руководитель – к.ф.-м.н., доцент Большакова Елена Игоревна

Дипломная работа посвящена созданию программных средств для автоматического извлечения информации (Information Extraction) из текстов на естественном языке. Information Extraction (IE) – быстро развивающееся прикладное направление в области компьютерной лингвистики, изучающее задачи извлечения определенной информации из ЕЯ-текстов – терминов и их отношений, именованных сущностей (имен, персоналий, географических названий) и событий, в которых участвуют выделенные объекты.

В дипломной работе используется метод, основанный на поиске и последующем извлечении из текста языковых конструкций, заданных специальными шаблонами распознавания, отражающими лингвистическую информацию. Для описания шаблонов распознавания используется язык лексико-синтаксических шаблонов LSPL [1], предоставляющий средства описания конструкций естественного языка для их автоматического выделения в тексте с учетом особенностей русского языка. В работе рассматривается расширение языка, позволяющее описывать шаблоны извлечения и шаблоны генерации, которые в совокупности с шаблоном разпознавания задают декларативные правила для извлечения информации из распознанной в тексте конструкции либо в виде нового шаблона распознавания, либо в виде текста, составленного из элементов распознанной конструкции.

Правила, содержащие шаблон извлечения, позволяют находить, преобразовывать и извлекать из текста нужные языковые конструкции, при этом шаблон извлечения задает элементы извлекаемой конструкции и операции, которые должны быть выполнены над ними. Эти элементы можно преобразовать в любую морфологическую форму – либо задавая конкретные значения их морфологических признаков или согласовывая одни элементы с другими. Эти возможности удобно использовать при решении задачи автоматизированного построения глоссария (перечня терминов документа с их определениями) – для этого в тексте распознаются определения терминов и извлекаются в канонической форме. Например, правило

```
Term = NG1 NG2<c=gen> "мы" "называем" NG3
       =text> #NG3 "-" "это" #NG1 NG2
```

где NG – это шаблон, описывающий именную группу, позволяет получить из распознанной фразы *направленное движение заряженных частиц мы называем электрическим током* определение термина *электрический ток – это направленное движение заряженных частиц*.

Правила, содержащие шаблон генерации, позволяют генерировать новые шаблоны из элементов (фрагментов) распознанной конструкции. Для этого шаблон генерации содержит ссылки на эти элементы, включая ссылки на слово в той форме, как оно встретилось в тексте, и ссылки на начальную форму слова. В шаблоне генерации возможны также ссылки на значение конкретного морфологического признака слова. Такие шаблоны необходимы при терминологическом анализе текста, который проводится в несколько этапов. Например, найдя все словосочетания вида *интеграл Римана*, *интеграл Лебега*, можно извлечь из них фамилии ученых (Риман и Лебег) и автоматически сгенерировать новые шаблоны распознавания, позволяющие затем автоматически распознавать в тексте различные термины с указанными фамилиями, в частности: *теорема Римана*, *сфера Римана*, *мера Лебега* и т.д. Все это позволяет сделать шаблон INT:

```
INT = N1<"интеграл"> N2<c=gen>
      =pattern> N1 N2<"$N2.b", c=gen>
```

В дипломной работе были разработаны процедуры, реализующие обработку рассмотренных выше шаблонов извлечения и шаблонов распознавания языка LSPL. Эти процедуры реализованы в виде дополнительных компонентов библиотеки LSPL [2] и могут быть интегрированы в различные приложения по обработке текста. Библиотека доступна для свободного использования, исходный код находится по адресу http://gitorious.org/lspl/lspl. Шаблоны извлечения и шаблоны генерации опробованы при создании нескольких приложений, в частности, системы автоматизированного построения глоссариев.

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 Международной конференции Диалог '2007 – М.: Издательский центр РГГУ, 2007, с. 70-75.
2. Большакова Е.И., Носков А.А. Программные средства анализа текста на основе лексико-синтаксических шаблонов языка LSPL // Программные системы и инструменты: Тематический сборник, № 11 / Под ред. Королева Л.Н. – М.: Изд. отдел факультета ВМиК МГУ имени М.В.Ломоносова; МАКС Пресс, 2010, с. 61-73.