УДК 519.688

СИСТЕМА ДЛЯ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ ИЗ ТЕКСТОВ НА БАЗЕ ЛЕКСИКО-СИНТАКСИЧЕСКИХ ШАБЛОНОВ

Е.И. Большакова (*bolsh@cs.msu.ru*)
К.М. Иванов (*ivanov.kir.m@yandex.ru*)
А.С. Сапин (*alesapin@gmail.com*)
Г.Ф. Шариков (*egos_@mail.ru*)

МГУ имени М.В. Ломоносова, Москва

Описывается текущее состояние программной системы с открытым кодом, разработанной и применяемой для построения различных приложений по извлечению информации из текстов на русском языке. Извлекаемая информация специфицируется в виде лексико-синтаксических шаблонов и правил языка LSPL.

Ключевые слова: извлечение информации из текстов, лексико-синтаксические шаблоны, лингвистические правила, язык LSPL, распознавание конструкций по шаблонам

Введение

Извлечение информации из текстов на естественном языке (ЕЯ) [Grishman, 2003] – одно из актуальных научных направлений, результаты которого используются во многих приложениях, включая реферирование и аннотирование текстов, извлечение знаний из специализированных текстов и др. К извлекаемой информации относят именованные сущности (персоналии, названия организаций и т.п.), их свойства, события, а также термины и понятия определенной предметной области и их связи.

Для упрощения разработки конкретных ЕЯ-приложений, в том числе по извлечению информации, применяются инструментальные системы, включающие стандартные модули анализа текста, средства сборки и отладки приложений, а также формальные языки для задания лингвистической информации о распознаваемых в текстах конструкциях. Лингвистическая информация задается обычно в форме специальных шаблонов и правил – с их помощью готовые программные модули анализа текста настраиваются для решения прикладной задачи.

К широко известным инструментальным системам относится GATE [GATE, 2016] с языком Jape для записи шаблонов и правил. Однако Jape не содержит лингвистической специфики, что требует дополнительной настройки на язык анализируемых текстов, весьма значительной для высокофлективных ЕЯ. Для более эффективной разработки приложений по извлечению информации из русскоязычных текстов были созданы системы со своими средствами задания шаблонов и правил: RCO Pattern Extractor [Ермаков и др., 2003], программный комплекс для языка LSPL [Большакова и др., 2010], процессор языка DSTL [Скатов и др., 2010], система [Томита-парсер, 2016]. Каждая из этих систем имеет свои особенности и ограничения, общим является применение частичного синтаксического анализа для извлечения из текста необходимых конструкций, их описание на формальном языке лингвистических шаблонов и правил. В целом, формальные языки этих систем сопоставимы по выразительным средствам задания лексических, графематических, морфосинтаксических особенностей извлекаемых конструкций.

В данной работе характеризуется текущее состояние программной системы для поиска и извлечения из ЕЯ-текстов конструкций по их описанию в виде *лексико-синтаксических шаблонов* языка LSPL [LSPL, 2016]. В последние годы в язык были введены новые средства, повышающие его выразительные возможности и упрощающие тем самым построение приложений. Рассматривается применение языка и системы для решения нескольких прикладных задач: извлечение информации из текстов финансовых обзоров, автоматическое построение глоссариев и предметных указателей для специализированных текстов.

1. Лексико-синтаксические шаблоны и правила

Язык LSPL [Большакова и др., 2010] создавался для декларативного описания лингвистических свойств конструкций русского языка, с целью автоматического их распознавания в системах обработки ЕЯ-текстов. Распознаваемая конструкция специфицируется в виде лексико-синтаксического шаблона, определяющего входящие в него слова с учетом их морфологических характеристик и условий грамматического согласования, типичного для многих выражений русского языка (в том числе именных словосочетаний: *битовый массив, шина адреса* и т.п.).

Основные встроенные возможности языка включают:

- конкретизацию для слов распознаваемой конструкции части речи, лексемы и морфологических характеристик (падеж, род, число и т.п.);
- задание имени шаблона и его параметров (из числа характеристик входящих в шаблон элементов-слов), что позволяет применять уже определенные шаблоны в качестве вспомогательных при описании шаблонов более сложных ЕЯ-выражений;

- указание условий грамматического согласования для слов-элементов шаблона, а также для вспомогательных шаблонов-элементов;
- опциональные, альтернативные, повторяющиеся элементы шаблона.

Набор взаимосвязанных шаблонов фактически задает КС-грамматику (расширенную условиями) распознаваемой языковой конструкции.

Приведем в качестве примера шаблон, описывающий однородные сочинительные конструкции вида *горы, яркое солнце и синее море* или *компьютер, ноутбук, планшет, а также другие устройства*:

AN { "," AN } ["и" AN | "," "а" "также" AN]

Метасимволы |, *{, }, [,]* означают соответственно альтернативные, повторяющиеся, опциональные элементы шаблона, в кавычках задаются строки, *AN* – вспомогательный шаблон: *AN = {A} N <A=N> (N)* , который определяет сочетание существительного (*N*) и нескольких согласованных с ним (<*A=N*>) прилагательных (*A*).

Поскольку в конкретных приложениях языка обычно требовалось некоторое преобразование распознанных по шаблону ЕЯ-выражений, в язык были введены дополнительные средства, позволившие задавать лингвистические правила вида:

шаблон_распознавания *=text>* шаблон_извлечения_текста

и шаблон_распознавания *=pattern>* синтезируемый_шаблон.

Новая возможность языка – шаблон извлечения текста (стоящий в правой части правила) дает возможность выделить составные части распознанной конструкции и сформировать из них нужную текстовую строку. К примеру, правило

A N <понятие> <A=N> =text> #A

позволяет извлечь в нормализованном виде (символ #), т.е. в словарной форме, все найденные прилагательные к существительному *понятие*.

В шаблоне извлечения кроме элементов распознанной конструкции и операции их нормализации можно устанавливать новые значения морфологических признаков этих элементов, а также применять к ним операцию грамматического согласования.

В правой части лингвистического правила может быть указан синтезируемый шаблон, т.е. способ построения нового шаблона из элементов распознанной конструкции. При этом кроме основных средств могут быть использованы ссылки (помечаемые знаком $) на элементы, использованные в левой части правила, и их морфологические признаки. К примеру, следующее правило служит для распознавания сочетаний из двух существительных, первое из которых конкретизировано (*понятие*), а второе – нет (*понятие решетки*, *понятие импульса* и т.п.) и построения шаблона для поиска именных сочетаний со вторым словом:

N1<понятие> N2<c=gen> =pattern> {A} N<$N2.b> <A=N>

В этом правиле используется ссылка на лемму второго слова (*$N2.b*), и согласно правилу будут сгенерированы шаблоны для всех найденных в тексте вариантов второго слова, например:

{A} N<решетка> <A=N > и *{A} N<импульс> <A=N>*

Синтезируемые шаблоны необходимы для поэтапного извлечения информации из текста, когда в нем сначала выделяются определенные конструкции, а затем их фрагменты образуют новый шаблон для продолжения поиска.

Введение в язык LSPL рассмотренных новых средств существенно упростило разработку приложений, требующих проведения сложных операций над текстом. Важной отличительной особенностью LSPL в сравнении с языками других инструментальных систем для извлечения информации из русскоязычных текстов является встроенная возможность синтеза новых шаблонов.

2. Функции и состав программной системы

Основной функцией системы [LSPL, 2016] является выделение в тексте и извлечение конструкций, согласно заданным шаблонам и правилам. При этом применяется определенная последовательность этапов обработки текста: токенизация (графематический анализ), морфологический анализ, распознавание конструкций по шаблонам на базе ранее разработанного метода [Носков, 2009], извлечение текста или генерация новых шаблонов. Система реализована на языке C++, ее исходный код является открытым (доступен по ссылке: https://github.com/cmc-msu-ai/lspl).

Основные программные компоненты системы:

- Центральный компонент, реализующий распознавание в тексте языковых конструкций по LSPL-шаблонам и их преобразование в извлекаемый текст или синтезируемый шаблон;
- Подключаемые модули графематического и морфологического анализа текста (в данный момент используются анализаторы [AOT, 2016]);
- Консольная утилита, реализующая обращение к центральному компоненту и вывод результатов работы в специальный XML-файл;
- Среда с графическим пользовательским интерфейсом для анализа текстов по шаблонам, предназначенная для лингвистов и/или специалистов по предметной области анализируемых текстов, которые участвуют в создании LSPL-шаблонов.

При создании приложений на базе LSPL в основном использовалась утилита, входные файлы которой должны содержать анализируемый текст и набор шаблонов и правил. Если правила включают правую часть (т.е. шаблоны извлечения текста и/или синтезируемые шаблоны), то утилита выдает в выходном файле результаты соответствующего преобразования найденных текстовых фрагментов (конструкций), т.е. извлеченный текст

или сгенерированные шаблоны. В ином случае утилита помещает в выходной файл только найденные текстовые фрагменты с сопутствующей информацией (морфологическими характеристиками слов фрагментов).

3. Визуальная среда анализа текстов

Построение конкретных приложений по извлечению информации из текстов предполагает разработку и отладку лингвистических шаблонов и правил. Входящая в состав системы визуальная среда поддерживает этот процесс, позволяя задавать различные LSPL-шаблоны конструкций, инициировать распознавание соответствующих конструкций и их извлечение, просматривать и анализировать полученные результаты.

Среда реализована на базе открытой версии библиотеки Qt, за счет чего достигается её кроссплатформенность (на данный момент для ОС семейств Linux и Windows). Заметим, что использование среды выгодно отличает систему LSPL от инструмента [Томита-парсер], в котором просмотр результатов анализа возможен только в отдельном файле.

Среда предоставляет следующие возможности:

- Загрузку и сохранение анализируемых текстов в любых кодировках;
- Создание и редактирование шаблонов с поддержкой истории и подсветкой синтаксиса;
- Загрузку шаблонов из файлов и их сохранение;
- Просмотр сообщений об ошибках, обнаруженных в шаблонах;
- Поиск и выделение в загруженном тексте конструкций по заданным шаблонам; возможность выгрузки результатов в XML-файл;
- Подсчет статистики выявленных конструкций;
- Просмотр морфологических характеристик слов текста;
- Возможность сохранения в файл (в формате XML) и последующей загрузки текущего контекста анализа (текст + шаблоны + результаты).

Пользовательский интерфейс состоит из трех основных, связанных между собой областей – см. Рис. 1.

В области анализируемого текста (левая верхняя часть экрана) найденные по шаблонам конструкции выделяются желтым цветом, и при наведении на них курсора мыши появляется всплывающая подсказка с морфологической информацией.

В области шаблонов (правая верхняя часть) расположен список всех шаблонов и правил, загруженных из файлов или введенных в среде; ниже расположено поле для ввода нового шаблона. Шаблоны и правила, применяемые для анализа текста в текущий момент, помечаются.

В нижней области, в виде таблицы выводится информация о результатах анализа, с возможностью фильтрации по примененным шаблонам. В первом столбце таблицы представлены выделенные

фрагменты (конструкции), во втором – результаты извлечения по примененному правилу, а в третьем – их морфологические параметры.

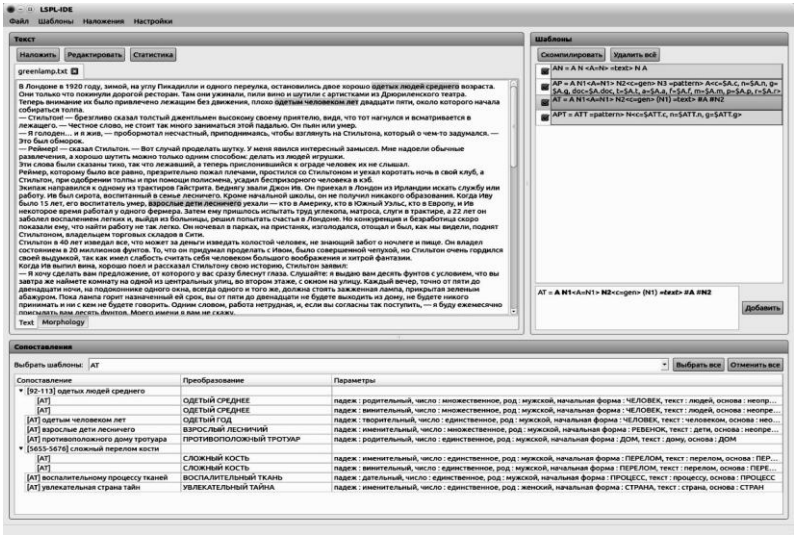

Рис. 1. Среда анализа текстов по шаблонам

4. Приложения, построенные на базе системы

С момента создания программной системы для языка LSPL было построено порядка десятка различных приложений, требующих извлечения и анализа информации из текстов. Наиболее крупным из них был комплекс процедур для автоматического терминологического анализа русскоязычных научно-технических текстов [Ефремова, 2013]. Были также приложения совсем другого типа, в частности, вопросно-ответные системы (вопросно-ответная система по теории элементарных чисел).

Рассмотрим три приложения, в которых для обработки текстов уже использовались лингвистические правила LSPL. К ним относится система для извлечения информации из текстов финансовых обзоров, выпускаемых аналитическими департаментами инвестиционных компаний и публикуемых в сети Интернет [Большакова и др., 2012]. Каждый из обрабатываемых текстов содержит упоминание о выпуске некоторой компанией финансовой отчетности за определенный временной период, что представляет собой извлекаемое из текста событие, например: *Вчера Автоваз подвел финансовые итоги за 3-й квартал 2012 года. Выручка компании выросла на 57 %, а себестоимость – на 40 %, в результате чего маржа по валовой прибыли составила 12.2 %...*

Из текстов интернет-обзоров извлекались такие атрибуты события:
– название компании, опубликовавшей отчетность (*Автоваз*);
– отчетный период (в приведенном фрагменте – *3-й квартал 2012 года*);
– изменение выручки компании (во фрагменте – *выросла на 57%).*

Составление необходимого набора LSPL-шаблонов происходило итеративно, путем последовательного их тестирования на заранее собранной коллекции текстов обзоров и уточнения – в итоге было получено более 90 шаблонов. Проведенные на их базе эксперименты показали довольно высокую эффективность извлечения: точность извлечения каждого атрибута события оказалась более 90%, а полнота в среднем превышала 75%.

Еще одним приложением, реализованным с использованием правил LSPL, была система автоматизированного построения глоссариев. Глоссарий специализированного документа в норме должен содержать все основные определяемые в нем термины, в форме упорядоченного по алфавиту списка глоссов – фраз вида: *Термин –Толкование*. Для выявления в тексте конструкций-определений терминов документа и последующего преобразования их в глоссы был модифицирован набор LSPL-шаблонов, разработанный ранее в рамках комплекса для терминологического анализа научно-технических текстов.

Приведём пример правила для построения глосса:

"под" Term <c=ins> ["обычно" | "здесь"] V1<пониматься, p=3, t=pres>
Defin<c=nom> <Defin.n = V1.n>
=text> Term<c=nom> "– " Defin<c=nom>

Здесь вспомогательный шаблон *Term* задает грамматическую структуру термина, *Defin* – шаблон определяющей термин фразы (толкование). Если это правило применить к тексту *...под экономическими ресурсами понимаются все природные, людские и произведенные человеком ресурсы, которые используются для производства товаров и услуг...*, то в результате распознавания и извлечения получим текст: *Экономические ресурсы – все природные, людские и произведенные человеком ресурсы, которые используются для производства товаров и услуг.*

Другим, близким по назначению, но более сложным разрабатываемым приложением является система поддержки построения предметного указателя для заданного текстового документа. Предметный указатель (*back-of-the-book-index*) представляет собой структурированный перечень обсуждаемых в документе ключевых терминов, понятий и объектов предметной области (ПО), с указанием страниц, на которых они встречаются в тексте. Основными проблемами автоматического построения указателей (весьма далекими до полного решения) являются:
– извлечение из текста терминов, понятий и названий объектов ПО;
– фильтрация и отбор наиболее важных (ключевых);

– выявление связей отобранных понятий (синонимии и ассоциативных);
– определение наиболее важных мест их употребления в документе.

Для решения этих подзадач используются различные лексико-синтаксические шаблоны языка LSPL, в том числе задающие извлечение терминов из типичных фраз-определений терминов, а также реализующие выявление терминологических синонимов, которые часто вводятся вместе с основным термином (например: *...будем называть определителем, или детерминантом матрицы*...). Кроме представительного набора шаблонов в рассматриваемом приложении применяется предложенная в [Ефремова, 2013] стратегия последовательного извлечения из текста терминологических словосочетаний с учетом разных типов шаблонов.

Заключение

В работе описана программная система, предназначенная для распознавания в текстах на русском языке конструкций по их формальному описанию в виде лексико-синтаксических шаблонов языка LSPL и извлечению из них нужной информации. LSPL-шаблоны и правила оказались достаточно гибким и мощным средством для разработки различных по характеру и сложности ЕЯ-приложений. Существенную роль при этом играют новая встроенная возможность спецификации извлекаемого теста и синтеза новых шаблонов, а также разработанная визуальная среда анализа текстов по шаблонам.

Опыт применения языка позволил выявить направления дальнейшего развития его выразительной мощности, к которым относится введение в язык логических операций, применяемых к условиям согласования и конкретизации морфологических характеристик. Более принципиальным является введение специальной операции-связки ~, обозначающей произвольный порядок элементов шаблона, что дает возможность компактно описывать такие конструкции, как глагол и его дополнение (которые могут стоять в произвольном порядке: *V ~ N*). Указанные средства пока отсутствуют во встроенных языках инструментальных систем для извлечения информации из текстов.

Список литературы

[GATE, 2016] General Architecture for Text Engineering [Электронный ресурс]. – Электрон. дан. – URL: http://www.gate.ac.uk/ (дата обращения: 09.07.2016)

[Grishman, 2003] Grishman R. Information extraction. In: The Oxford Handbook of Comput. Linguistics. Mitkov R. (ed.). Oxford University Press, 2003.

[LSPL, 2016] Lexico-Syntactic Pattern Language: Описание проекта [Электронный ресурс]. – Электр. дан. – URL: http://www.lspl.ru/ (дата обращения: 09.07.2016)

[**AOT**] AOT: Автоматическая обработка текстов [Электронный ресурс]. – Электрон. дан. – URL: http://www.aot.ru/ (дата обращения: 09.07.2016)

[Большакова и др., 2010] 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.: МАКС Пресс, 2010.

[Большакова и др., 2012] Большакова Е.И., Жеребцова Ю.А. Эксперименты по извлечению информации из аналитических текстов финансовых обзоров // Информационные системы для научных исследований: Сборник научных статей. Труды XV Всерос. объединенной конф. "Интернет и современное общество". Санкт-Петербург, 2012.

[Ермаков и др., 2003] Ермаков А.Е., Плешко В.В., Митюнин В.А. RCO Pattern Extractor: компонент выделения особых объектов в тексте // Информатизация и информационная безопасность правоохранительных органов: XI Международная научная конференция. Сборник трудов – Москва, 2003.

[Ефремова, 2013] Ефремова Н.Э. Методы и программные средства извлечения терминологической информации из научно-технических текстов: дис. ...канд. физ.-мат. наук: 05.13.11. – М.: 2013.

[Носков, 2009] Носков А.А. Метод выделения в тексте конструкций по их лексико-синтаксическим шаблонам // Сборник статей молодых ученых фак-та ВМК МГУ – М.: МАКС Пресс, 2009, Выпуск 6.

[Скатов и др., 2010] Скатов Д.С., Вдовина Н.А. и др. Язык описания правил в системе лексического анализа ЕЯ-текстов Dictascope Tokenizer // Комп. лингвистика и интеллектуальные технологии: По материалам Междунар. конф. «Диалог» (Бекасово, 26-30 мая 2010 г.) Вып. 9 (16). – М.: Изд-во РГГУ, 2010.

[Томита-парсер, 2016] Яндекс: Томита-парсер. [Электрон. ресурс]. – Электрон. дан. – URL: https://tech.yandex.ru/tomita/ (дата обращения: 09.07.2016)

A SYSTEM FOR INFORMATION EXTRACTION FROM TEXTS BASED ON LEXICO-SYNTACTIC PATTERNS

E.I. Bolshakova (*bolsh@cs.msu.ru*)
K.M. Ivanov (*ivanov.kir.m@yandex.ru*)
A.S. Sapin (*alesapin@gmail.com*)
G.F. Sharikov (*egos_@mail.ru*)

Moscow State Lomonosov University, Moscow

Current state of an open source system developed to create various applications for information extraction from Russian texts is described. Information to be extracted is specified by lexico-syntactic patterns and rules of LSPL language.

Keywords: information extraction from texts, lexico-syntactic patterns, linguistic rules, LSPL language, phrase recognition based on patterns