Москвина Анна Денисовна

Санкт-Петербургский государственный университет (СПбГУ),
Санкт-Петербург, Россия
moskvina.anya@gmail.com

Орлова Дарья Витальевна

Санкт-Петербургский государственный университет (СПбГУ),
Санкт-Петербург, Россия
frenezo@mail.ru

РАЗРАБОТКА СИНТАКСИЧЕСКОГО АНАЛИЗАТОРА ДЛЯ РУССКОГО ЯЗЫКА С ПОМОЩЬЮ КАТЕГОРИАЛЬНОЙ ГРАММАТИКИ В NLTK И СИНТАКСИЧЕСКИХ ПРАВИЛ АОТ

Ключевые слова: автоматическая обработка текстов, синтаксический анализатор, русский язык, NLTK.

Целью работы является создание синтаксического анализатора для русского языка с использованием инструментов NLTK для Python. Анализ проводится на основании разработанной нами формальной грамматики и использует морфологическую разметку, принятую в системе PyMorphy2. При создании грамматики мы адаптировали под нашу систему правила, описанные в проекте АОТ. В статье обсуждаются возможности синтаксического компонента NLTK, преимущества выбранной категориальной грамматики, особенности разработанных правил и алгоритм работы программы.

Moskvina Anna

Saint Petersburg State University (SPbSU),
St. Petersburg, Russia

Orlova Darja

Saint Petersburg State University (SPbSU),
St. Petersburg, Russia

DEVELOPMENT OF PARSER FOR RUSSIAN LANGUAGE USING FEATURE-BASED GRAMMAR IN NLTK AND SYNTACTIC RULES

Keywords: natural language processing, syntactic analysis, Russian, NLTK.

Our work is aimed at the development of the syntactic parser for Russian based on NLTK toolkit for Python. The parser is based on the formal grammar we have developed and the tagset accepted in PyMorphy2 morphological tagger. We have adjusted the rules described in the AOT project for our purposes. The article describes how we can analyze sentence structure with NLTK, the advantages of feature-based grammar for Russian and how our grammar works with the most important syntactic groups occurring in Russian texts.

Введение

Среди проблем автоматической обработки информации в компьютерной лингвистике синтаксический анализ играет значительную роль, так как применяется в широкой области задач: автоматическая коррекция текста, машинный перевод, извлечение сущностей и т. д. Являясь наиболее сложным этапом досемантической обработки текста, проблема автоматического синтаксического анализа русского языка до сих пор не нашла общепринятого решения. На данный момент существующие парсеры, или синтаксические анализаторы, для русского языка чаще всего оказываются коммерческими или узкоспециализированными. Нам представляется, что путь к решению проблемы — в развитии открытых некоммерческих лингвистических платформ.

Данная статья состоит из 2 глав. Первая глава посвящена уже готовым компонентам синтаксического анализатора: синтаксису в NLTK и категориальным грамматикам, а также правилам, описанным в проекте АОТ. Во второй главе в общих чертах описывается процесс разработки анализатора: особенности правил, программная реализация и перспективы работы. Более детально изучить механизм работы ядра анализатора можно в другой нашей статье [Москвина и др., 2016].

Основные компоненты синтаксического анализатора

Синтаксис в NLTK

Одно из первых мест среди открытого программного обеспечения занимает NLTK (Natural Language Toolkit) [Bird et al.], набор библиотек и инструментов для языка программирования Python, предназначенный для основных процедур автоматической обработки текстов. В NLTK уже встроены морфологические и синтаксические анализаторы для работы, например, с английским языком, а также в нём есть всё необходимое для самостоятельной разработки своей формальной грамматики или анализатора. Подключая к встроенному в NLTK парсеру свою грамматику, мы можем получить на выходе структуру составляющих, которые отражают то, как слова и последовательности слов сочетаются, формируя синтаксические группы.

Категориальные грамматики

Мы работаем с таким видом формальной грамматики, как категориальная грамматика, основанная на признаках категорий

(feature-based grammar). Это один из видов порождающих грамматик Хомского, содержащих в себе нетерминальные элементы, терминальные элементы и правила порождения. В категории, которые заложены в правилах порождения, мы можем вписать признаки, то есть грамматические характеристики слова, и с их помощью мы можем регулировать согласование в одной группе. Другими словами, каждой категории приписывается некоторый набор изменяемых параметров, информация о значении которых используется при выделении синтаксических групп. Это позволяет явно указывать морфологические особенности компонентов сочетания, напр.:

NP[CASE=?c, GENDER=?g, NUMBER=?n] -> Adj[CASE=?c, GENDER=?g, NUMBER=?n] Noun[CASE=?c, GENDER=?g, NUMBER=?n]

Приведенное правило описывает объединение прилагательного и существительного в именную группу. Здесь мы работаем с тремя категориями. С помощью переменных ?c, ?g, и ?n мы обозначаем согласование по, соответственно, падежу, роду и числу прилагательного и существительного и сохраняем те же значения признаков у получившейся именной группы.

Такой подход представляется нам выгодным для русского языка, обладающего развитой морфологией.

Правила выделения синтаксических групп

При разработке правил категориальной грамматики мы опирались на синтаксические группы, описанные в документации синтаксического модуля AOT [AOT: Синтаксический анализ]. Проект «AOT» (Автоматическая обработка текста) с 2002 года занимается многоуровневым анализом текста и выкладывает свои данные в открытый доступ. Синтаксический компонент на данный момент не представлен в виде отдельной работающей программы, однако с его описанием можно ознакомиться на официальном сайте. Синтаксический анализ в AOT основывается на морфологической информации и направлен на выявление синтаксических групп в одной клаузе.

Взяв за основу некоторые синтаксические группы проекта AOT, мы разработали правила выделения синтаксических групп. Для того, чтобы правила работали корректно, нам необходима была морфологическая информация, и для этого мы сначала применяем морфоанализатор PyMorphy2 [Korobov, 2015], также написанный на языке Python.

Разработка синтаксического анализатора

Некоторые особенности нашей грамматики

Напомним, что для разработки синтаксического анализатора мы, во-первых, воспользовались уже готовыми синтаксическими группами проекта АОТ, во-вторых, «перевели» эти правила на язык категориальной грамматики, заложенной в NLTK, и, в-третьих, для синтаксического анализа обратились к выходным данным морфологического анализа, произведённого PyMorphy2, которые также вложили в правила категориальной грамматики.

В правилах применяется рекурсия, т. е. группа из правой части правила повторяется в левой. Мы можем вводить в правила некоторый параметр, имеющий булево значение. Так мы сохраняем информацию о некоторой подкатегории, не меняя имя самой группы и обеспечивая рекурсию. Сохранение информации о подкатегории необходимо для введения некоторых ограничений на объединение. В качестве примера рассмотрим правило объединения переходного глагола с объектом в несколько сокращенном виде:

VP[+objt, <…>] -> VP[-objt,<…>] NP[CASE=accs]

Такая запись обозначает, что правило может примениться для одной глагольной группы только один раз. В правой части правила параметр objt имеет значение «ложь», но после срабатывания правила параметр приобретает значение истины и предупреждает рекурсивное срабатывание. Таким образом, в словосочетании *вижу лапу кота*, глагольная группа *вижу лапу* уже не сможет присоединить к себе слово *кота* в качестве объекта, хотя оно и следует линейно за группой и стоит в винительном падеже.

Программная реализация

Разработанная нами категориальная грамматика для русского языка в виде списка правил записывается в файл. Пользователю предлагается ввести предложение или словосочетание для разбора. Каждое слово обрабатывается морфоанализатором PyMorphy2, а полученные морфологические параметры словоформ представляются в виде терминальных элементов в категориальной грамматике NLTK, напр.:

NOUN[CASE=gent, GENDER=femn, NUMBER=sing, PERS=3, NF=u ‘рука’] -> ‘руки’

Перспективы

На сегодняшний день разрабатываемый парсер успешно справляется с разбором простых предложений. В будущем мы планируем увеличить количество правил, а также ввести механизм выбора более вероятного разбора предложения, в основе которого будет лежать концепция силы связей между членами предложения, принятая в теории конструктивного синтаксиса Н. Ю. Шведовой [Русская грамматика, 1980].

По итогам проведенного исследования можно утверждать, что идея создания синтаксического анализатора на основе категориальной грамматики является состоятельной. На данный момент мы продолжаем работу, чтобы перейти к тестированию парсера на русскоязычном корпусе текстов и оценке результатов его разборов.

ЛИТЕРАТУРА

АОТ: Синтаксический анализ. Построение дерева зависимостей всего предложения. URL: http://www.aot.ru/docs/synan.html

Москвина А. Д., Орлова Д., Паничева П. В., Митрофанова О. А. Разработка ядра синтаксического анализатора для русского языка на основе библиотек NLTK // Труды XIX Международной объединенной научной конференции «Интернет и современное общество». СПб., 2016.

Русская грамматика. Т. 2: Синтаксис / Н. Ю. Шведова (гл. ред.). М.: Наука, 1980.

Bird S., Klein E., Loper E. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. URL: http://www.nltk.org/book/

Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages. Analysis of Images, Social Networks and Texts // 4th International Conference, AIST 2015. Yekaterinburg, Russia, April 9–11, 2015.