4. Колби, Р. Энциклопедия технических индикаторов рынка. – Л.: Альпина Паблишер, 2013. – 837 с.
5. Акелис, С. Технический анализ от А до Я. – Л.: Омега-Л, 2010. – 376 с.

УДК 004.912

РАЗРАБОТКА СИСТЕМЫ ПОСТРОЕНИЯ ДЕРЕВА СИНТАКСИЧЕСКОГО ПОДЧИНЕНИЯ ДЛЯ АНАЛИЗА РУССКОЯЗЫЧНЫХ ТЕКСТОВ

Лёвкин А.В., Тарасова И.А.
Донецкий национальный технический университет
кафедра искусственного интеллекта и системного анализа
E-mail: a.lewckin2010@yandex.ru

Аннотация:
Лёвкин А.В., Тарасова И.А. Разработка системы построения дерева синтаксического подчинения для анализа русскоязычных текстов. *Проведено исследование методов анализа текста, способов описания синтаксической структуры. Разработана система построения дерева синтаксического подчинения для анализа текстов русского языка, позволяющая выполнять обработку и анализ тексов больших объемов.*

Annotation:
Lovkin A.V., Tarasova I.A. Development of a system for building a tree of syntactic subordination for the analysis of Russian texts. *A study of text analysis methods, ways to describe the syntactic structure. A system for constructing a tree of syntactic analysis for the analysis of texts in Russian has been developed.*

Постановка задачи

Автоматический анализ естественно-языковых текстов является востребованной технологией, которая находит применение в текстовых процессорах (например, Microsoft Word, OpenOffice.org Writer) и поисковых системах, системах реферирования, системах классификации и кластеризации текстов и, наконец, в системах поиска дубликатов в текстах.

Задача синтаксического анализа является одной из сложных задач компьютерной лингвистики. Исследования в этой области начались еще в 1960х годах. Были созданы различные системы, которые позволяли проводить синтаксический анализ предложений на естественном языке. Эти разработки существенно продвинули теорию и практику синтаксического анализа, однако, полученные программные реализации не обеспечивают высокое качество анализа.

Цель статьи – минимизация времени на обработку и анализ русскоязычных текстов больших объемов за счет разработки системы построения дерева синтаксического подчинения.

Для достижения поставленной цели были решены следующие задачи:

- проведено исследование методов анализа текста;
- проведено исследование способов описания синтаксической структуры;
- разработана система построения дерева синтаксического подчинения для анализа текстов русского языка.

Исследования

Синтаксический анализ текста включает в себя следующие этапы:

- разбиение текста на базовые элементы;
- морфологический анализ;

- графематический анализ;
- построение дерева синтаксического подчинения.

Синтаксический анализ предназначен для выделения лексических и нелексических единиц текста в целях дальнейшей их обработки синтаксическим и семантическим компонентами лингвистического процессора [1,3].

Этап разбиения текста на базовые элементы предполагает выделение в тексте последовательностей символов одного алфавита (алфавит символов кириллицы, алфавит символов латиницы, алфавит знаков препинания, алфавит разделителей, алфавит для записи числовых значений, алфавит скобочек и кавычек, все остальные символы). Такие последовательности символов называются базовыми элементами.

На этапе морфологической разметки текста выполняется морфологический анализ написаний базовых элементов, состоящих из символов латиницы или кириллицы.

Этап графематического анализа позволяет интерпретировать базовые элементы и последовательности базовых элементов как нелексические единицы с определенной семантической нагрузкой (e-mail, URL, дата, имя файла и подобные им) или лексические единства, выступающие в предложении как единое целое [2].

Задачей синтаксического анализа является построение синтаксической структуры входного предложения (осуществление разбора предложения) на основе морфологической информации о словоформах и синтаксических правил объединения слов и словосочетаний. Синтаксическая структура отражает синтаксические связи, существующие между словами в предложении. Ее получение начинается с построения всевозможных связей между словами, которые в последующем отсеиваются на основе локальных и глобальных «фильтров». Конкретный вид структуры определяется выбранной системой синтаксических отношений (ССИО) [4,5].

Существует несколько способов описания синтаксической структуры, но два из них – система составляющих и дерево зависимостей – являются наиболее эффективными.

Для описания синтаксической структуры с помощью деревьев зависимостей (деревьев синтаксического подчинения) необходимо определить на множестве X точек цепочки x бинарное отношение таким образом, чтобы граф был (ориентированным) деревом с корнем. Всякое такое отношение называется отношением (синтаксического) подчинения, а соответствующее дерево – деревом (синтаксического) подчинения. При графическом изображении дерева подчинения, обычно располагают точки или, так называемые, цепочки на горизонтальной прямой и проводят стрелки сверху от нее. Тем самым получается «естественное» дерево подчинения, изображённое на рисунке 1.

Рисунок 1 – Дерево синтаксического подчинения (дерево зависимостей)

Корнем дерева подчинения принято считать сказуемое, т.к. оно представляет собой организующий элемент предложения.

Дуги дерева подчинения часто снабжаются метками, указывающими типы представляемых этими дугами синтаксических связей. Системы составляющих и деревья зависимостей характеризуют синтаксическую структуру предложения в разных аспектах. С помощью систем составляющих описываются в явном виде словосочетания, но игнорируется ориентация связей.

Одно предложение может допускать несколько различных "естественных" систем составляющих (деревьев подчинения). Чаще всего это бывает в тех случаях, когда смысл предложения можно понимать по-разному, и разные системы составляющих (деревья подчинения) отвечают разным толкованиям смысла [6].

Иногда для представления синтаксической структуры предложения используют смешанное представление, называемое обобщенной синтаксической структурой (ОСС). ОСС выражает, как и дерево зависимостей, ориентацию связей, но, в отличие от дерева зависимостей, ОСС снабжена информацией о словосочетаниях, образованных группами членов предложения (группой подлежащего, группой сказуемого, группой дополнений, группой обстоятельств и т.п.).

Реализация системы построения дерева синтаксического подчинения для анализа текстов русского языка осуществлена в Visual Studio 2010 в приложении MFC. Система позволяет анализировать заданный текст, передаваемый из файла, разбивать его на сегменты и выводить в виде дерева синтаксического подчинения.

На рисунке 2 представлен вид окна, которое выводится на экран при запуске. Для дальнейшей работы системы пользователю необходимо задать файл с текстом.

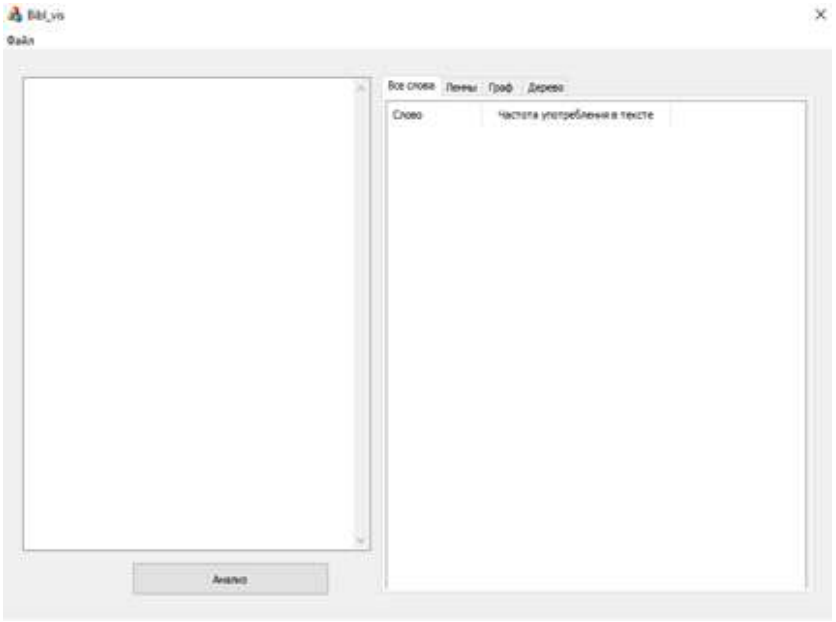

Рисунок 2 – Главное меню

На рисунке 3 представлен результат разбиения исходного текста на сегменты.

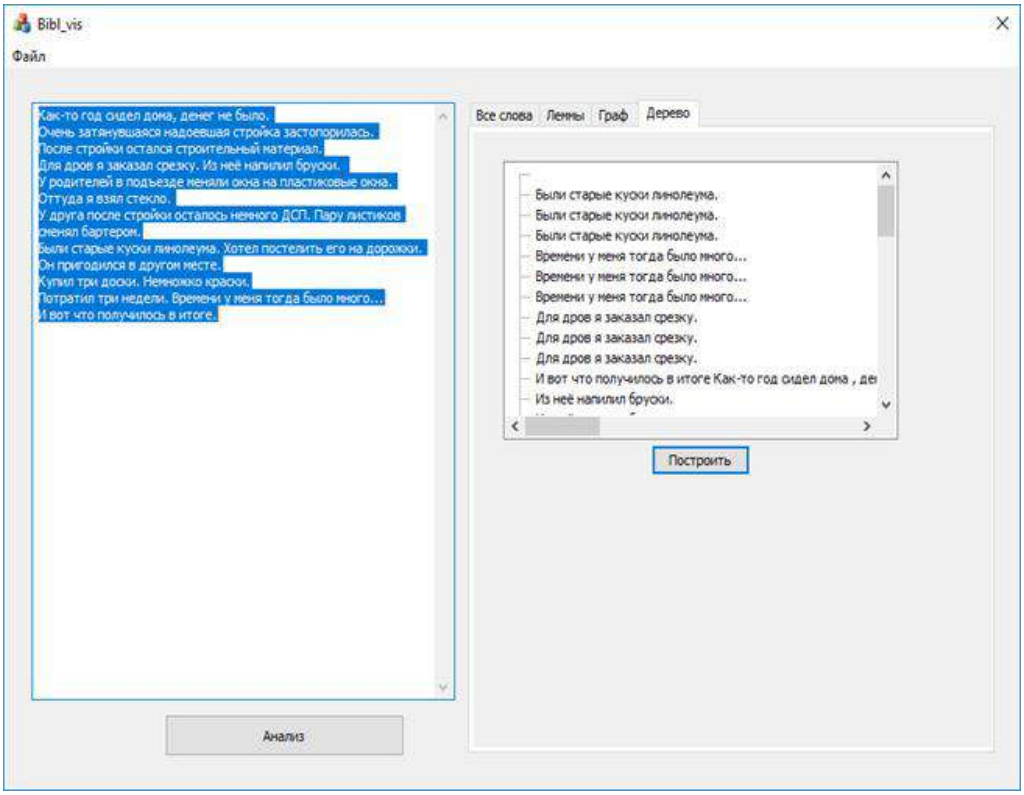

Рисунок 3 – Результат разбиения текста на сегменты

На рисунке 4 представлен результат описания синтаксической структуры с помощью дерева синтаксического подчинения.

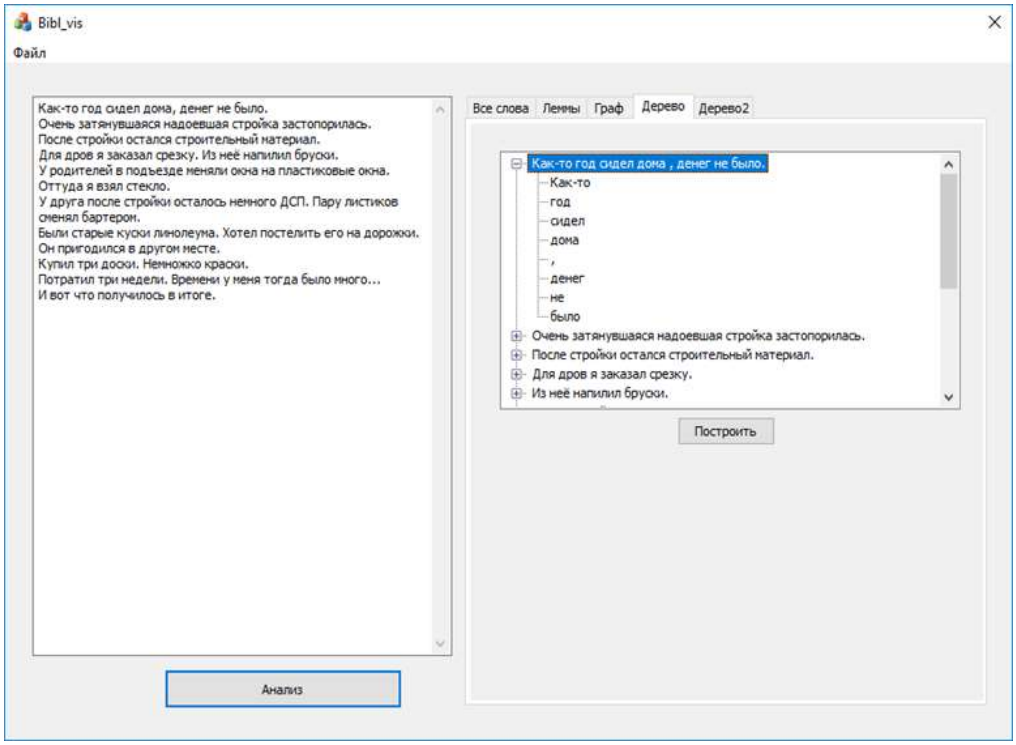

Рисунок 4 – Результат построения дерева синтаксического подчинения

Выводы

Разработанная система синтаксической обработки текста, описанная в данной работе, от существующих отличается тем, что морфологическая разметка выполняется до графематического анализа, после чего строится дерево синтаксического подчинения, позволяющее сохранить ориентацию связей. Это дает возможность дальнейшего использования дерева синтаксического подчинения в других задачах анализа текста.

Литература

1. Автоматическая Обработка Текста [Электронный ресурс].– Режим доступа: http://www.aot.ru/technology.html

2. Сокирко, А.В. Семантические словари в автоматической обработке текста / А.В. Сокирко // Диссертация на соискание ученой степени кандидата технических наук. – М.: МГПИИЯ. 2001. – 108 с. [Электронный ресурс]. – Режим доступа – http://www.aot.ru/docs/graphan.html

3. Дорохина, Г.В. Модуль морфологического анализа слов русского языка / А.П. Павлюкова, Г.В. Дорохина // Искусственный интеллект. Донецк: – ІПШІ: Наука і освіта, 2004. – № 3. – С. 636-642.

4. Дорохина, Г.В. Модуль морфологического анализа без словаря слов русского языка / Г.В. Дорохина, В.Ю. Трунов, Е.В. Шилова // Искусственный интеллект. – №2. – 2010. – С.32-36.

5. Бондаренко Е. А. Принципы автоматической обработки естественно-языковых текстов: валентностный подход / Е.А. Бондаренко, О.А. Каплина // Искусственный интеллект. – 2013. – №1. – С. 80-90.

6. Galina V. Dorokhina The Algorithm of Syntactic Analysis Based on Grammatical Rules // Искусственный интеллект. - №4, 2014. – С. 169 – 179.