РАЗРАБОТКА ИНФОРМАЦИОННОЙ СИСТЕМЫ ДЛЯ ПОСТРОЕНИЯ И АНАЛИЗА КОРПУСА РУССКИХ ПУБЛИЦИСТИЧЕСКИХ ТЕКСТОВ XIX ВЕКА СМАЛТ

К. А. Кулаков, А. В. Седов
Петрозаводский государственный университет
Петрозаводск
kulakov@cs.karelia.ru, sedov_a@mail.ru

Одним из ключевых методов лингвистического исследования является корпусный метод. Данный метод позволяет систематизировать и быстро получать информацию о текстах, объединённых по определённым правилам и имеющим схожую разметку. В Петрозаводском государственном университете работы по компьютерной обработке текстов ведутся с 1995 года в рамках проекта «Статистические методы анализа литературных текстов» (СМАЛТ). Атрибуция текста проводится с использованием разборов, полученных с помощью программного комплекса СМАЛТ, однако в процессе создания корпусов текстов пришлось столкнутся с созданием нескольких разметок для одно и того же набора текстов. Цель работы заключается в обеспечения комфортной работы по атрибуции текстов, построению русскоязычного корпуса и проведению лингвистических исследований в современных информационных средах. В работе представлено описание архитектуры разрабатываемого программного комплекса, его функциональные возможности и особенности реализации.

Ключевые слова: корпусный метод, синтаксический разбор, семантический разбор, разметка текста.

DEVELOPMENT OF AN INFORMATION SYSTEM FOR THE CONSTRUCTION AND ANALYSIS OF THE CORPUS OF RUSSIAN JOURNALISTIC TEXTS OF THE XIX CENTURY SMALT

K. A. Kulakov, A. V. Sedov
Petrozavodsk state university
Petrozavodsk

One of the key methods of linguistic research is the corpus method. This method allows you to systematize and quickly receive information about texts combined according to certain rules and having similar markup. In Petrozavodsk State University, works on computerized text processing have been carried out since 1995 in the framework of the project Statistical Methods for the Analysis of Literary Texts (SMALT). Attribution of the text is carried out with the use of parsing, obtained using the software complex SMALT, however, the process of creating text boxes had to deal with the creation of several markup for the same set of texts. The purpose of the work is to ensure comfortable work on the attribution of texts, the construction of the Russian-language corpus and the conduct of linguistic studies in modern information environments. The paper presents a description of the architecture of the developed software system, its functionality and implementation features.

Key words: corpus method, syntactic parsing, semantic parsing, text markup.

Современные лингвистические исследования не обходятся без использования информационных систем с различными алгоритмами. Одним из ключевых методов исследования является корпусный метод. Данный метод позволяет систематизировать и быстро получать информацию

о текстах, объединённых по определённым правилам и имеющим схожу разметку. Это позволяет повысить эффективность работы специалистов — филологов, а так же позволяет независимым исследователям проверять достоверность полученных данных и проводить проверку результатов исследований. Одним из ярких примеров создания русскоязычных корпусов является Национальный корпус русского языка (http://www.ruscorpora.ru) [1, 2].

Несмотря на развитие компьютерной техники, значительному распространению сети интернет (интернет пришёл в каждый дом), созданию систем типа «умный дом», «голосовой помощник» и других, задача семантического анализа до сих пор не является решенной. Для решения данной задачи необходим комплексный анализ текстов, что влечёт за собой работу с огромным массивом данных, количество которых растёт с каждым годом. Это касается не только современных текстов, но и текстов предыдущих эпох (с каждым годом оцифровывается множество различных текстов, восстанавливаются рукописи, проводятся исследования работ некоторых авторов, проводятся открытия, которые меняют авторов некоторых произведений, что заставляет пересматривать взгляды на творчество некоторых писателей).

В Петрозаводском государственном университете (ПетрГУ) работы по компьютерной обработке текстов ведутся с 1995 года. Их результатом явилась разработка программного комплекса «Статистические методы анализа литературных текстов» (ПК «СМАЛТ») [1], имеющего в своей основе базу данных текстов, состоящую из публицистических статей разной тематической направленности из петербургских журналов «Время», «Эпоха», «Современник», «Гражданин» «Светоч», «Молва», «Библиотека для чтения», «Заря» XIX века [3] в оригинальной орфографии.

Атрибуция текста производится при помощи разборов, созданных в ПК «СМАЛТ». Имеется различный набор правил, построенный как с использованием методик Г. Хетсо [4], так и с использованием различных эвристик. Но выбранный способ атрибуции не является определяющим, а лишь используется как подсказка для специалистов — филологов, которые с учётом полученных данных принимают решение. Большое разнообразие признаков и их комбинаций делает задачу атрибуции текстов трудоемкой. В процессе создания корпусов текстов пришлось столкнутся с созданием нескольких разметок для одно и того же набора текстов [4–6]. К сожалению преобразования между разметками не было взаимооднозначным, что приводило в необходимости «двойной работы» при анализе текстов. На данный момент существуют различные программы для предварительной разметки и обработки текстов, однако большинство из них ориентированно на современные тексты и современную графику, что по сути делает их использование невозможным для произведений в графике XIX века. Так же зачастую набор признаков, выделяемый данными программами, не совместим с наборами признаков, построенным специалистами —филологами для ПК «СМАЛТ». При этом в ПК «СМАЛТ» были реализованы некоторые алгоритмы для начальной разметки, но они до сих пор находятся в зачаточном состоянии [5]. Но с момента создания и внедрения данных алгоритмов прошло достаточно большое время, в течение которого появились проверенные временем алгоритмы, позволяющие проводить изначальную разметку с высокой точностью, что теоретически должно помочь специалистам избежать дополнительных ошибок при разборе и ускорить процесс создания корпуса (перейти от скучной монотонной работы по разметке и перейти уже непосредственно к основной части лингвистического исследования). Также различная скорость разбора произведений различными исследователями иногда приводила к торможению

общего процесса. Отсутствие достаточно удобного механизма взаимодействия в процессе разбора (разбор производился в режиме offline), заставляло некоторых исследователей ждать общего собрания для обсуждения деталей. Также привлечение новых специалистов требует проверки качества их работы, что было проблематично ввиду отсутствия информации о том, кто производил разбор.

Программный комплекс СМАЛТ разрабатывался несколькими поколениями студентов, аспирантов и сотрудников ПетрГУ, что привело к появлению большого числа сложно-совместимых компонент и модулей. Некоторые модули имели различную направленность: некоторые использовали оригинальное написание, некоторые требовали современного представление тех же текстов. При этом, большая часть программного комплекса разрабатывалась с использованием ныне устаревших версий языков программирования, технологий хранения данных для не поддерживаемых в настоящее время операционных систем.

Несмотря на достаточно большое число различных корпусов текстов русского языка, а также средств создания своих корпусов, актуальной является задача обеспечения комфортной работы по атрибуции текстов, построению русскоязычного корпуса и проведению лингвистических исследований в современных информационных средах. Использование клиент-серверной архитектуры позволяет выполнять совместные работы с синхронизацией данных на сервере, что позволяет избежать проблем с потерей данных частью сообщества исследователей, а также позволяет другим участникам оперативно получать новую информацию. Ключевыми особенностями предметной области являются большой объем разнородной информации (например, грамматические, синтаксические, семантические разборы) и большое количество операций с текстом (например, ручная разметка частей текста, грамматический разбор слова в тексте, разбор всего текста, анализ разборов текстов).

В качестве хранилища выбрана документо-ориентированная база данных MongoDB (https://www.mongodb.com/). Она позволяет хранить данные в виде структурированных документов (объектов), что позволит работать с произвольным содержимым, позволяя создавать объекты, удобные для использования как специалистами — филологами, так и техническим персоналом.

Серверная часть (Backend) реализована на базе микрофреймворка Slim3 (https://www.slimframework.com/) (PHP). Микрофреймворк позволяет быстро и легко организовать REST интерфейс (https://ru.wikipedia.org/wiki/REST) для загрузки и сохранения данных, а также выдачи необходимых результатов по запросу.

Трудоемкие задачи (например, автоматический разбор текстов) вынесены за пределы сервера и реализованы в виде обработчиков для платформы Gearman (http://gearman.org/). Платформа Gearman позволяет выполнять распределенные вычисления с использованием очереди задач и может быть масштабирована в зависимости от потребностей.

Клиентская часть реализована на базе библиотеки построения интерфейсов пользователя React JS (https://reactjs.org/). Библиотека позволяет реализовать функционально-ориентированный подход, когда интерфейс пользователя подстраивается под выполняемые задачи и текущие состояния.

Для хранения и управления данными на клиентской стороне используется библиотека Redux (https://redux.js.org/). Она позволяет организовать единое хранилище данных для всех компонент интерфейса пользователя, выполнять обновление состояний компонент интерфейса и последующую перерисовку.

Планируется преобразование текущей архитектуры ПК и построение модульной архитектуры, где за каждую часть работы ПК будет отвечать соответствующий набор компонент. Данное изменение добавит гибкости в использовании ПК и позволит конфигурировать систему под необходимые задачи.

Планируется выделить следующие модули:

- Модуль грамматического анализа с ручным и/или автоматическим разбором;
- Модуль синтаксического анализа с ручным и/или автоматическим разбором;
- Модуль семантического анализа с ручным и/или автоматическим разбором;
- Модуль атрибуции текста с механизмами шаблонов и подсказок;
- Модуль исследований на базе подготовленного корпуса языка с подключением различных алгоритмов;
- Модуль публикации информации для доступа сторонним пользователям;
- Модуль идентификации пользователя;
- Модуль совместной работы;
- Модуль статистической обработки и представления статистической информации.

В ходе разработки ПК планируется выполнить импорт разборов текстов прошлых лет в различных наборах атрибутов для последующего анализа и систематизации. Также подготовлен перечень из более чем 400 текстов различных авторов XIX века, требующих разбора и атрибуции. Таким образом, задача реализации программного комплекса для разбора и атрибуции текстов весьма актуальна.

Поддержка исследований. Исследование выполнено при финансовой поддержке РФФИ (Отделение гуманитарных и общественных наук), проект «Проблема атрибуции анонимных и псевдонимных статей в журналах „Время“, „Эпоха“ и еженедельнике „Гражданин“ (No 18-012-90026).

Библиографический список

1. Лингвистические корпусы: монография / А. А. Котов, З. И. Минеева, А. А. Рогов и др. — Петрозаводск: Изд-воПетрГУ, 2014. — 140 с. ISBN 978-5-8021-2066-8.
2. Рогов А. А. Программный комплекс „СМАЛТ“ / А. А. Рогов, Г. Б. Гурин, А. А. Котов, Ю. В. Сидоров, Т. Г. Суровцова// Труды 10-й Всероссийской научной конференции „Электронные библиотеки: перспективные методы и технологии, электронные коллекции“ — RCDL’2008, Дубна, Россия, 2008.
3. Захаров В. Н. Программная система поддержки атрибуции текстов статей Ф. М. Достоевского / В. Н. Захаров, А. А. Леонтьев, А. А. Рогов, Ю. В. Сидоров // Труды Петрозаводского государственного университета: Сер. Прикладная математика и информатика. Вып. 9. — Петрозаводск, 2000.
4. Хетсо Г. Принадлежность Достоевскому: к вопросу об атрибуции Ф. М. Достоевскому анонимных статей в журналах Время и Эпоха // Oslo: Solum Forlag A. S., 1986. — 82 с.
5. Котов А. А. Информационная система для создания размеченных корпусов малой размерности / А. А. Котов, М. Ю. Некрасов, А. В. Седов, А. А. Рогов // Ученые записки Петрозаводского государственного университета. — 2012. — №. 8–1. — С. 108–112.
6. Рогов А. А., Сидоров Ю. В., Суровцова Т. Г. Математические методы атрибуции литературных текстов небольшого объема // Материалы XIII Всероссийской конференции» Математические методы в распознавании образов».-М.: МАКС Пресс. — 2007. — С. 525–528