ИНФОРМАЦИОННО-АНАЛИТИЧЕСКАЯ СИСТЕМА «СМАЛТ»[1]

THE INFORMATION-ANALYTICAL SYSTEM "SMALT"

***Рогов А.А**. (rogov@psu.karelia.ru), **Сидоров Ю.В**. (sidorov@psu.karelia.ru), **Солопова А.И.**, **Суровцова Т.Г** (tata@nlcom.onego.ru), Петрозаводский государственный университет*

В данной работе представлена информационная система «СМАЛТ». Основная ее цель заключается в сборе, централизованном хранении литературных произведений вместе с их грамматической и синтаксической структурами, а также в статистической обработке и анализе этих структур для выявления закономерностей.

Введение

Мысль «поверить алгеброй гармонию» языка давно овладела умами исследователей, хотя сложность применения математических методов в этой области не вызывает сомнений. Трудность, прежде всего, состоит в неоднозначности языковых единиц, функционирование которых в тексте определяется взаимодействием самых разнообразных лингвистических и экстралингвистических факторов. При попытке перевести языковую абстракцию в абстракцию математическую исследователь неизбежно наталкивается на то, что «исследуемое явление оказывается практически неразложимым на дискретные единицы, обнаружение которых – необходимое условие успешной статистической обработки» [1]. Тем не менее необходимо искать методы, использование которых позволит максимально отразить те тончайшие нюансы семантико-грамматических связей, сочетание которых дает тот или иной эффект и составляет индивидуальные особенности творческой манеры автора и художественное своеобразие произведения.

Кроме того, сама проблема разбора литературных текстов по различным параметрам (морфологическим, синтаксическим, лексическим) является достаточно трудоемкой. Как правило, для решения различных прикладных задач, связанных с разбором текстов по определенным параметрам, до недавнего времени исследователи редко когда подвергали ручной обработке литературное произведение полностью, ограничиваясь лишь выборками, которые в совокупности покрывали не более 10% текста. Естественно, что результаты таких исследований могли содержать невольно допущенные искажения и ошибки и могли подвергаться критике. Еще одним немаловажным фактором является то, что материал, накопленный в результате обработки одной группой исследователей, зачастую не мог быть использован другими, тогда как подобная возможность могла бы привести к пересмотру некоторых ранее полученных результатов и, вероятно, появлению новых интересных для научного сообщества фактов. Эти и ряд других сложностей в ручной обработке текстов привели к возникновению идеи о необходимости создания единой информационной системы автоматизированного сбора, обработки и хранения информации о литературных текстах.

В Петрозаводском государственном университете работы по компьютерной обработке текстов велись с 1995 года, и их результатом явилась разработка информационной системы «Статистические методы анализа литературных текстов» (ИС «СМАЛТ»), имеющей в своей основе базу данных текстов (публицистических статей разной жанровой и тематической направленности из петербургских журналов «Время» и «Эпоха» за период с 1861-1865 гг.). Известно, что, являясь редактором этих изданий, Ф.М. Достоевский публиковал собственные статьи анонимно или под другим именем. В.В. Виноградов, в статьях, посвященных описанию стиля Достоевского-публициста, назвал его «мастером языка и бытового фельетона» и отметил, что эта сторона его творчества мало исследована [2].

Общее описание системы «СМАЛТ»

ИС состоит из двух основных блоков: *функционального блока*, предназначенного для морфологического и синтаксического анализа текстов, пополнения БД литературных произведений, а также внесения исправлений; и *аналитического блока*, состоящего из модулей, реализующих разнообразные методики статистического анализа текстов.

[1] Проект поддержан грантами РГНФ № 02-04-12015в (руководитель А. А. Рогов) и № 05-04-12418в (руководитель А. А. Рогов)

В качестве исходного источника данных для клиентского приложения используется текстовый файл в кодировке Unicode, что позволяет избежать проблем, связанных с использованием в отдельных текстах символов, специфичных как для отдельных языков, так и для орфографии разных периодов одного языка. Обработка текстов в информационной системе производится в несколько этапов. На первом шаге выполняется автоматизированное разбиение исходного текста на лексические единицы, среди которых выделяются часть (или раздел), абзац, предложение, слово. Разбиение осуществляется на основе аппарата регулярных выражений. Шаблон разбиения можно изменять от текста к тексту при этом он сохраняется вместе с данными о разбиении. На втором этапе осуществляется автоматическая обработка текста и его морфологический разбор. При морфологическом разборе для отдельных частей речи выделяется до 20 морфологических признаков. На базе построенного морфологического разбора производится третья стадия обработки текста – синтаксический анализ. На этой стадии для каждого предложения исходного текста выделяются в среднем около 15 признаков. После осуществления обработки входного текста, ее результаты помещаются в централизованное хранилище (репозиторий текстов, готовых для статистического анализа).

На следующем этапе пользователь может выполнять операции по анализу текстов, находящихся в репозитории как с использованием клиентского программного обеспечения, так и частично через WEB, используя предоставляемый web-узлом интерфейс. Кроме этого пользователям ИС СМАЛТ предоставляется возможность внесения изменений и поправок в опубликованные данные. Таким образом, можно просмотреть одни и те же данные в редакции различных специалистов, а также сравнить результаты, получаемые при статистической обработке различных редакций.

Особенности грамматического разбора

Выделение грамматической единицы в модуле морфологического анализа СМАЛТ производится автоматически по признаку ограничения ее с обеих сторон пробелами. Таким образом, атрибутированию последовательно подлежит всякое выделенное по этому принципу слово. В этом случае мы сталкиваемся с проблемой характеристики морфологем, выражающих свое грамматическое значение аналитически или в виде идиоматического сочетания. За рамками точного анализа могут остаться сложные формы глагольного времени, степеней сравнения, составные наречные, союзные, предложные, междометные сочетания, огромное количество разного вида фразеологизмов, части которых и вовсе невозможно рассматривать в отдельности. Поэтому при разработке системы морфологического анализа был избран путь, который предполагает атрибуцию с использованием расширенного арсенала грамматических категорий. Таким образом, для каждой выделенной единицы предусмотрена наиболее близкая ей по контексту семантико-грамматическая характеристика, что позволяет точнее определить ее статус в морфологической иерархии текста. Т.е. в атрибуционной системе для элементов составных конструкций предусмотрены такие значения, как: *«часть составного союза*» (с дальнейшей характеристикой союза), «*часть производного предлога*» (с указанием его типа и значения), «*часть сложной формы времени*», «*часть фразеологизма*» и т.п. В процессе дальнейшей синхронизации данных разделенные элементы составных морфологем будут представлены как единая морфологическая единица текста. Кроме того, учитывается наличие в тексте неязыковых символов, цитат, афоризмов, иностранных слов и пр.

Решая задачу выявления субъективного начала в объективном, СМАЛТ предлагает расширенную систему атрибуции разного рода переходных и переносных значений внутри грамматических парадигм, а также многозначности слов в общей парадигме частей речи. Подобный подход может значительно изменить общую статистическую картину и дает возможность выделить более тонкие нюансы смысловых и лексико-грамматических отношений в тексте, «особые тенденции внутренней динамики слов» [2].

Одним из наиболее распространенных подходов к автоматизации морфологической разметки текстов на русском языке на сегодняшний день является использование словаря А.А. Зализняка. Однако в связи с тем, что ИС «СМАЛТ» изначально была рассчитана на анализ мультиязыковой текстовой информации, использование словаря, ориентированного на морфологию отдельного выделенного языка оказалось невозможным. Кроме этого в случае обработки текстов в старой транскрипции большая часть словоформ будет отличаться от современного варианта и появится необходимость значительного расширения используемого словаря. В связи с вышеперечисленными факторами выбор был сделан в пользу собственного словаря, состоящего из словоформ и связанных с ними морфологических характеристик (для отдельных словоформ количеством до 20). Длительность процесса заполнения словаря оказалась приемлемой, так настоящее время объем словаря морфологических единиц русского языка достиг 50000 словоформ. Данный словарь демонстрирует многообразие типов морфологической омонимии и представляет богатый лексико-грамматический материал.

Однако с ростом объема словаря увеличивается и процент неоднозначностей в морфологической, а как следствие и синтаксической разметках текста. На первоначальном этапе для частичного снятия омонимии

использовались эвристические методы, базировавшиеся на ряде правил согласования по времени, роду, числу и падежу. В дальнейшем данный подход оказался трудно масштабируемым и практически неприменимым в условиях анализа текстов на разных языках. Более перспективным представился подход, основанный на реализации универсального модуля предсинтаксической обработки [3]. Метод, в основе которого лежит разрешение морфологической омонимии на базе наращиваемого набора правил (или ярлыков и условий) легко может быть адаптирован к работе в условиях мультиязыковой системы. Авторами выделяются 58 типов омонимии. Одним из базовых компонентов предлагаемой авторами системы универсального модуля предсинтаксиса является словарь диагностических ситуаций.

Особенности синтаксического разбора

На этапе преформатирования текста в ИС СМАЛТ производится автоматическая разбивка текстового материала на дискретные синтаксические единицы, выделяемые по пунтуационному признаку конца предложения и формально равные ему. В модуле синтаксического разбора СМАЛТ заложен максимально возможный набор основных значений для характеристики структурно-семантических и модально-интонационных типов предложений. Атрибутированию не подвергаются цитаты и афоризмы (в статистике они фигурируют с соответствующей пометой).

Идея отразить (хотя бы в некоторой степени) в статистическом исследовании особенности структурного построения текста, а также своеобразие синтаксической организации представленных к анализу материалов побудили создателей СМАЛТ включить в атрибутирующий комплекс параметр: «*особый характер связи в контексте*» со значениями: 1) «*нет*» (если предложение автосемантично), 2) «*часть сложного синтаксического целого*», 3) «*присоединение*» (если это единственный присоединяемый элемент) и некоторые другие. Так реализуется попытка «ранжировать» единицы текста применительно к их структуре, объему, текстообразующей роли и степени самостоятельности. Таким образом, статистическая картина наряду с другими показателями продемонстрирует функционирование предикативных единиц с точки зрения их автосемантичности или синсемантичности в структуре текста. Следует уточнить, что автосемантия (смысловая достаточность) и синсемантия (смысловая недостаточность) понимается здесь в двух аспектах: в структурном и смысловом. Группа синсемантичных предложений будет представлять сложное синтаксическое целое.

Что же касается присоединения, то в СМАЛТ принято четкое разграничение этого понятия на разных уровнях структурных построений. Если это осложняющая синтагма в рамках предложения, то она квалифицируется как «*присоединительная конструкция*» в параметрах осложнений простого предложения. Если присоединительная конструкция выдвигается в парцеллят и выделяется в тексте как формально самостоятельное предложение, то ему в параметре «*особый характер связи в контексте*» будет присвоено значение: «*присоединение*». Если же присоединяемое предложение является компонентом более крупного структурного образования, то оно в этом же параметре определяется как «*часть сложного синтаксического целого*». Таким образом, в статистике проявляются и способы выражения структурно-синтаксической динамики на уровне разных компонентов текста.

В связи с этим работа по выделению подлежащих атрибуции синтаксических единиц в условиях автоматизированного анализа делает особенно важным внимание к особенностям их пунктуационного оформления. Автоматически членение текста на единицы анализа производится, как указывалось выше, по формальным показателям конца предложения, в качестве которых приняты точка, вопросительный и восклицательный знаки. Однако два последних не всегда следует рассматривать только в этом качестве. На этапе преформатирования текста существует возможность изменить расположение специальных символов, выделяющих синтаксическую единицу для последующего анализа. В зависимости от поставленных задач, исследователь может, таким образом, атрибутировать всю конструкцию как единое целое или рассматривать ее части как отдельные предложения, определяя их структурно-семантический статус в параметрах, о которых говорилось выше. Это позволит в какой-то мере проследить особенности структурной организации текста и, возможно, сыграет свою роль в выявлении авторской индивидуальности.

Особенности пунктуации исследуемых нами текстов требует пристальное внимание к некоторым деталям. Так, например, интересным представляется различие в способах оформления предложений с прямой речью. Некоторые наблюдения позволяют отметить следующую особенность: если прямая речь в тексте действительно представляет собой слова другого лица (вымышленного или настоящего), то это подчеркивается разграничением планов выражения автора и этого лица, что отражается в пунктуации. Если же «прямая речь» представляет лишь предполагаемую автором мысль предполагаемого же оппонента, мы наблюдаем как бы слияние их субъектно-речевых планов в единый полифоничный «комплекс мыслей» одного человека, т.е. автора. Чтобы подчеркнуть эти различия, СМАЛТ может атрибутировать подобные типы предложений по-разному: один – как **кон-**

струкцию с прямой речью, а другой – как **сложную синтаксическую конструкцию,** маркируя их таким образом по этому признаку.

Несмотря на то что, на данный момент, синтаксический разбор в системе «СМАЛТ» частично выполняется вручную специалистом филологом, предполагается реализация парсера нотации HPSG и разработка грамматики для автоматизации синтаксиса русского языка в рамках этой теории.

Модуль статистической обработки

Целью модуля статистической обработки текстов является выделение скрытых закономерностей внутри текстов, анализ возможных зависимостей между текстами, поиск параметров, не изменяющихся в пределах диапазона текстов одного автора, одной стилистической направленности, одной эпохи, одного языка и т.д. В качестве параметров (признаков) текста может выступать любой набор морфологических или синтаксических характеристик исследуемых объектов. Набор признаков для исследования пользователь определяет самостоятельно при помощи любых комбинаций признаков из общего комплекта. Модуль статистической обработки текстов, находящихся в репозитории, представляет собой набор независимых библиотек, содержащих функции, позволяющие выполнять:

- разбиение заданного множества текстов на кластеры с заданием типа кластера;
- проверку гипотезы сходства текстов на основе критериев однородности распределения частотных характеристик параметров текстов.

Среди реализованных математических методов выделим следующие:

- методы проверки гипотез (критерий Стьюдента, непараметрический критерий Колмогорова);
- ряд критериев на согласованность с заданным распределением (Хи-квадрат, Колмогорова-Смирнова);
- методы кластерного анализа (иерархического кластерного анализа (ближней связи, средней связи Кинга), метод корреляционных плеяд, экстремальной группировки признаков и т.д.);
- методы факторного и компонентного анализов.

Кроме этого реализована простейшая нейронная сеть Хэмминга, выполняющая функцию отнесения входного текста к одному из эталонных классов.

При статистической обработке текстов в ИС есть возможность экспорта данных во внешние приложения и статистические пакеты для специфической обработки.

Полученные результаты

ИС «СМАЛТ» в основной своей части была разработана в сентябре 2002 г. Реализованный WEB – интерфейс доступа к единой БД литературных текстов находится по адресу http://smalt.karelia.ru. В результате эксплуатации на данный момент единая БД текстов насчитывает 246 разборов текстов из литературно-критических журналов «Эпоха», «Время», «Светоч» и т. д. На базе ИС «СМАЛТ» был выполнен ряд исследований по поиску авторского инварианта, некоторые результаты которых опубликованы в [4-7]. Среди них выделим исследования связанные с использованием метода «сильного графа» [4, 6], и на основе статистических методов, используемых Г. Хетсо в книге [8]. Отличием от исследования Г. Хетсо являлось использование большого количества похожих текстов других авторов в орфографии XIX века. Несмотря на отдельные отличия, результаты его исследования в основном подтвердились. Главным выводом, к которому пришли авторы, является то, что не существует единого инварианта для задачи атрибуции литературных тексов. Для решения каждой конкретной задачи надо использовать свой оригинальный набор признаков, который подбирается экспериментально.

Список литературы

1. Ахманова О.С., Натан Л.Н., Полторацкий А.И., Фатющенко Ф.И. О принципах и методах лингво-стилистического исследования // М.: 1966.
2. Виноградов В.В. История русского литературного языка // М.: Наука, 1978.
3. Кобзарева Т.Ю., Афанасьев Р.Н. Универсальный модуль предсинтаксического анализа омонимии частей речи в РЯ на основе словаря диагностических ситуаций // Труды Международного семинара Диалог'2002 по компьютерной лингвистике и ее приложениям. – http://www.dialog-21.ru.
4. Захаров В.Н., Рогов А.А., Сидоров Ю.В. Проблема грамматического инварианта Достоевского и атрибуция анонимных и псевдонимных статей в журналах «Время» и «Эпоха» (1861-1865) // Труды и материалы Международного конгресса «Русский язык: исторические судьбы и современность» (13-16 марта 2001 г.). М.: МГУ, 2001. С.404-405.

5. Рогов А.А., Сидоров Ю.В., Король А.В. Автоматизированная система обработки и анализа литературных текстов “СМАЛТ” // Труды и материалы II-го Международного конгресса исследователей русского языка “Русский язык: исторические судьбы и современность” (18-21 марта 2004 года). М.: МГУ, 2004. С.485-486.
6. Суровцева Т.Г. Анализ синтаксического разбора публицистических произведений Ф.М. Достоевского // Труды ПетрГУ, сер. Прикладная математика и информатика, вып.12, Петрозаводск: ПетрГУ, 2006. С. 72-82.
7. Суровцева Т.Г. Использование метода «сильного графа» при анализе синтаксического разбора публицистических произведений Ф.М. Достоевского // Труды ПетрГУ, сер. Прикладная математика и информатика, вып.12, Петрозаводск: ПетрГУ, 2006. С.83-91.
8. Хетсо Г. Принадлежность Достоевскому: к вопросу об атрибуции Ф.М. Достоевскому анонимных статей в журналах “Время” и “Эпоха” // SOLUM FORLAG A.S.: OSLO 1986.