Аналитический обзор программных средств предобработки корпусов текста.

Темников Роман Геннадьевич
Студент 1го курса магистратуры
Московского технологического университета
г. Москва
E-mail: zector23@yandex.ru

Компьютерная обработка текстов, в первую очередь текстов, созданных на флективных языках, основывается на морфологическом и синтаксическом анализе синтагм, предложений в соответствии с правилами формальной грамматики. Работа программ опирается на статистическую основу — корпус текстов, которые предварительно аннотированы разработчиками и использованы для «обучения» программы, а также алгоритмическое индексирование той или иной словарной базы, обычно — словаря, которого снабжен морфологическим модификатором.

Russian Morphological Dictionary работает с входным ASCII-текстом. Используется морфологический словарь А.Зализняка, включающий 120.000 слов. Реализована на SWI-Prolog для Windows. Программа быстро и с опорой на указанный словарь определяет грамматические признаки слов. При обращении к текстам социолектной принадлежности — это может обеспечить доказательную атрибуцию морфов, используемых в речи пользователей социальных сетей.

Mystem— это компактный, очень быстрый и бесплатный морфологический парсер русскоязычных текстов, реализованный также на основе словаря А.Зализняка. Доступны для загрузки версии для Windows и Linux. Работает как консольное приложение и имеет различные режимы представления результатов. Отмечается сложность установки программы и введения нужных параметров исследования.

SDK Pullenti выделение именованных сущностей (разметка по частям речи) из неструктурированных русскоязычных. Типы сущностей: персоны, организации, даты, страны, указы и др. Выделение сущностей основаны на правилах.

TextAnalyst 2.0 произведен научно-производственным инновационным центром «МикроСистемы» как инструмент анализа символьных текстов. Позволяет построить семантическую сеть понятий, выделенных в обрабатываемом тексте, со ссылками на контекст.

Galaktika-ZOOM представляет собой автоматизированную систему поиска и аналитической обработки информации. Это мощный инструмент анализа и обработки текста (Text Mining), позволяющий извлекать необходимые сведения из огромного объема данных. При обработке запроса формирует еще и информационный портрет объекта — список значимых для полученной по запросу выборки слов и словосочетаний, которые и следует уточнить.

АОТ (автоматическая обработка текста) общее название инструментов обработки текста на естественном языке, разработанных при создании системы автоматического перевода ДИАЛИНГ. Пакет состоит из компонентов — лингвистических процессоров, которые друг за другом обрабатывают входной текст. Вход одного процессора является выходом другого. Среди предлагаемых продуктов:

1. модуль графематического анализа текста;
2. компоненты морфологического анализа для русск., нем. и англ.яз.;
3. модуль автоматического уничтожения омонимии;
4. модуль семантического анализа текста;
5. система лингвистического поиска (конкорданс);

AskNet- семантические вопросно-ответные поисковые системы AskNet и инструментарий разработчика, реализующий полный лингвистический анализ текстов на русском и английском языках.

Таблица 1- сравнение программных средств.

Программы	Язык	Реализация	Анализ
Russian Morphological Dictionary	Русский	Windows	Синтаксический/ морфологический
Mystem	Русский	Windows и Linux	Морфологический
SDK Pullenti	Русский/ Украинский/ Английский	Windows	Выделение сущностей/ семантический/ морфологический
TextAnalyst 2.0	Русский/ Английский	Windows 95 и выше	Семантическая сеть понятий
Galaktika-ZOOM	Русский	Windows	Графематический/ морфологический
AOT	Русский/ Английский/ Немецкий	Windows и Linux	Графематический/ морфологический/ семантический.
AskNet	Русский/ Английский	Windows	Графематический/ морфологический/ синтаксический/ семантический.

Проведена классификация информационных продуктов, что позволило определить область применения для каждого из рассмотренных информационных продуктов. Хочется оотметить SDK Pullenti в качестве решения предобработки корпусов текста из-за способности обрабатывать, по сравнению с остальными, большие объёмы данных, а также бесплатного распространения.

Безусловно, никакая программная обработка текста не может заменить собой анализ, который может осуществить человек — особенно эксперт в той или иной области. Однако программы, о которых здесь идет речь, позволяют специалисту прийти к заключениям о тенденциях, потратив на проведение исследования меньшее количество времени.

Список литературы:

1. Логичев С. В. (2002) Каталог лингвистических программ и ресурсов в Сети [электронный ресурс] URL: http://rvb.ru/soft/catalogue/catalogue.html
2. Баранов А.Н. (2007) Введение в прикладную лингвистику.
3. Гарабик Р., Захаров В.П. (2006) Параллельный русско-словацкий корпус // Труды международной конференции «Корпусная лингвистика — 2006».