Аналитический обзор программных средств предобработки корпусов текста.

Р.Г. Темников,
студент;
рук. П.Г. Круг (МИРЭА, МГУПИ, МИТХТ, 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, каждый элемент словника, которого снабжен морфологическим модификатором (модификаторами). Широко используется вероятностный подход. Существенными задачами такого анализа является машинная обработка значительных объемов информации и обобщенное представление ее основного смысла в сжатой форме, вычленение смысловых доминант и тематической структуры, определение формальных характеристик стиля и жанра.

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. Кроме того, эти программы позволяют апробировать гипотезы на большем объеме материала и с большей долей уверенности в объективности полученных данных. Именно с этих позиций и будут рассмотрены имеющиеся сегодня программы обработки русскоязычных текстов.

Первая группа компьютерных программ предназначена для синтаксического и морфологического анализа русскоязычных текстов. Грамматический срез — один из важнейших при формировании целостного представления о системе языка, так что эти программы могут быть полезны в нашем исследовании.

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. С другой стороны, необходимо иметь в виду проблему ограниченности словника словаря А.Зализняка, в котором отсутствуют имена собственные, некоторые неологизмы последнего времени, сравнительные формы вроде постарше, наречия вида по-детски, многие сложные слова, пишущиеся через дефис, многие наречия на -о и —е. Соответственно, прогнозируются затруднения при определении грамматической принадлежности новых для системы литературного языка слов.

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. В общем, программа Mystem производит морфологический анализ литературного нормативного текста на русском языке. Для слов, отсутствующих в словаре,

порождаются гипотезы на основании частотности суффиксов. Отмечается сложность установки программы и введения нужных параметров исследования.

SDK Pullenti выделение именованных сущностей (разметка по частям речи) из неструктурированных русскоязычных текстов (Named Entity Recognition for Russian Language). Типы сущностей: персоны, организации, даты, страны, указы и др. Выделение сущностей основаны на правилах. Некоторые типы сущностей могут отождествляться с записями внешних словарей, если таковые есть (например, готовыми списками сотрудников или организаций). В ограниченном объёме система работает с украинскими текстами. Удобно для использования в системах, разрабатываемых на .NET. Для Mac и Linux-систем работает на платформе Mono.

Вторая группа программ автоматической обработки текстов объединяет продукты, которые позволяют прийти к обобщенному представлению о частоте выявленных лексических единиц, об их группировке в текстах, а также дают основания для исследования семантических процессов в изучаемых речевых продуктах.

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. Имеется возможность смыслового поиска фрагментов текста с учетом скрытых в тексте смысловых связей со словами запроса. Позволяет анализировать текст путем построения иерархического дерева тем/подтем, затрагиваемых в тексте. Также имеется возможность реферирования текста.

Основные возможности:

1. анализ содержания текста с автоматическим формированием семантической сети с гиперссылками — получение смыслового портрета текста в терминах основных понятий и их смысловых связей;

2. анализ содержания текста с автоматическим формированием тематического древа с гиперссылками — выявление семантической структуры текста в виде иерархии тем и подтем;

3. смысловой поиск с учетом скрытых смысловых связей слов запроса со словами текста;

4. автоматическое реферирование текста — формирование его смыслового портрета в терминах наиболее информативных фраз;

5. кластеризация информации — анализ распределения материала текстов по тематическим классам;

6. автоматическая индексация текста с преобразованием в гипертекст;

7. ранжирование всех видов информации о семантике текста по «степени значимости» с возможностью варьирования детальности ее исследования;

8. автоматическое формирование полнотекстовой базы знаний с гипертекстовой структурой и возможностями ассоциативного доступа к информации.

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 «Галактика ZOOM», кроме списка документов, где содержится информация по тому объекту, который ищет пользователь, формирует еще и информационный портрет объекта — список значимых для полученной по запросу выборки слов и словосочетаний, которые и следует уточнить.

При работе с информационным портретом пользователь может получить общее

представление об объекте, уточнять запрос по отдельным словам, составляющим информационный портрет объекта, отсекать лишнюю информацию, определять связи между отдельными словами, составляющими информационный портрет, позволяет получать качественный информационный результат в кратчайшие сроки.

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);
6. различные тезаурусы и словники.

В третьей группе программных продуктов собраны те системы, которые позволяют собирать данные, необходимые для определения стилевой принадлежности текстов, а также степени оригинальности текстов и/или приверженности авторов текстов той или иной стилистической манере.

Свежий взгляд- это DOS-утилита, реализующая стилистическую проверку русскоязычных текстов. Программа отыскивает в тексте места, где фонетически и морфологически схожие слова расположены в непосредственной близости, что порождает паронимы (например, «минимально возможное количество информации, которое можно...».

Технологии поиска и анализа текстовой информации — это сайт, на котором представлены разработки известной компании Гарант-Парк-Интернет. Среди представленных технологий:

1. анализ и классификация текстов, автоматическое реферирование;
2. различные варианты поиска текста;
3. морфологический, синтаксический и семантический анализ текста;
4. средства навигации по большим массивам текстов.

Безусловно, автоматическое реферирование не может быть целью при углубленном филологическом анализе, однако использование этой программы позволит делать некоторые выводы относительно тем, волнующих профессиональные сообщества, т.к. автоматическая обработка текстов даст возможность получить факты относительно очень большого объема контентов.

Худломер связан с задачей автоматической классификации стиля русскоязычных текстов. Автором были собраны и проанализированы 4 корпуса текстов, взятых из русской сети. Сюда вошли художественные произведения, публицистика, научные статьи и протоколы диалогов через ICQ и IRC. В результате были получены эмпирические кривые распределения длин слов в текстах, в зависимости от стиля. Эти кривые используются в качестве эталонов при классификации.

Программа классифицирует стиль входного текста как: разговорная речи, художественная литература, газетная статья или научная статья. Представляется, что использование этой программы позволит сделать наблюдения над стилистической принадлежностью исследуемых

текстов вне зависимости от обсуждаемых тем, в то время как при чтении текста человеком обсуждаемая тема часто играет решающую роль при определении функционального стиля.

AskNet- семантические вопросно-ответные поисковые системы AskNet и инструментарий разработчика, реализующий полный лингвистический анализ текстов на русском и английском языках. Модули лингвистического анализа включают в себя морфологию (словарную и бессловарную), синтаксис, семантику (включая толково-комбинаторные словари).

Имеется модуль семантической рубрикации текстов. Программные продукты представлены коробочными версиями корпоративной, сайтовой и персональной поисковой системы. Вопросно-ответный поиск по Интернету реализован на базе метапоисковой системы www.asknet.ru. Разрабатывается аналитическая поисковая система AQUA, позволяющая находить семантические ответы на основе автоматического обобщения системой текстовой информации и проведения логического вывода. Программы и SDK распространяются на коммерческой основе. Уровни лингвистического анализа: графематический, морфологический, синтаксический, семантический.

Таблица 1- сравнение программных средств.

Программы	Язык	Число слов	Реализация	Анализ
Russian Morphological Dictionary	Русский	120.000	Windows	Синтаксический/ морфологический
Mystem	Русский	120.000	Windows и Linux	Морфологический
SDK Pullenti	Русский/ Украинский/ Английский	200.000	Windows	Выделение сущностей/ семантический/ морфологический
TextAnalyst 2.0	Русский/ Английский	100.000	Windows 95 и выше	Семантическая сеть понятий
Galaktika-ZOOM	Русский	90.000	Windows	Графематический/ морфологический
AOT	Русский/ Английский/ Немецкий	161.000	Windows и Linux	Графематический/ морфологический/ семантический.
Свежий взгляд	Русский	80.000	Windows	Стилистический
Худломер	Русский	100.000	Windows	Стилистический
AskNet	Русский/ Английский	150.000	Windows	Графематический/ морфологический/ синтаксический/ семантический.

Таким образом выбор программного продукта должен быть обусловлен тем, какими методами предобработки текста он располагает. Пользователь должен выбрать нужный продукт в зависимости от того, на каком языке он хочет обработать текст и какой анализ желает использовать.

Проведена классификация информационных продуктов, что позволило определить область применения для каждого из рассмотренных информационных продуктов.

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».