

АНАЛИЗ ТЕКСТА С ПОЗИЦИЙ ВОСПРИЯТИЯ ЕГО КАК СТАТИСТИЧЕСКОЙ СТРУКТУРЫ

Основу рукописи любого издания составляет текст. Чаще всего авторы представляют в редакцию завершённое литературное произведение в электронной форме (.doc, .txt, .rtf). Помимо традиционного содержания термина «текст» его можно воспринимать и как «набор символов, закодированных специальным образом для его дальнейшей обработки».

Преобразование морфологической структуры текста в статистическую показало, что статистическая структура всех текстов, созданных человеком, построена по единым правилам и неизменна. Как и в любом научном методе, статистическую структуру текста исследуют, рассматривают, разбирают, считывают и пр., устанавливая количественные закономерности взаимодействия множеств в некоем статистическом пространстве.

В компьютерные технологии обработки текстового материала входят: языковой, стилистический, семантический, лингвистический, сравнительный, сопоставительный анализы, а также терминологический, категориальный, сленговый, пробельный, контент- и интент-анализ, казуальный, метод предпочтений, рейтинговый, дискурсивный, смысловой и др. Для классификации используют поверхностные, синтаксические, лексические, ссылочные и дискурсивные признаки.

Компьютерный анализ текста на основе лексикографических признаков реализуется в текстовых редакторах для ввода и коррекции программ, данных и запросов к системе, а также в программах контроля орфографии и разметки, сегментации, лемматизации, получения статистики словоупотреблений, загрузки, поиска и коррекции данных и др. Результатом такого анализа являются частотные словари, конкордансы, автоматические моно- и многоязычные словари.

Анализ текста на основе лингвистического обеспечения направлен на автоматическое индексирование входных

документов, составление поисковых предписаний по тексту запросов и ведение словарей системы. Синтаксический и семантический анализ лежит в основе системы автоматического перевода и систем человеко-машинного общения. Данные формы анализа опираются на формальный аппарат, основанный на законах математической и вычислительной лингвистики. Грамматический разбор элементов текстового материала осуществляется парсерами [1].

Выделяют также специализированные средства обработки данных, среди которых инструменты реализации контент-анализа и пакеты статистической обработки: STADIA, SYSTAT, STATGRAPHICSPLUS, SPSS, STATA, STATISTIKA, JMR, NCSS, MINITAB 14, PRISM, Статэксперт, Эвриста, R, EpiInfo, PSPP, SOFA и др. [2]. Стандартный набор статистических операций установлен в программе Microsoft Excel (макрос-дополнение XLSTAT-Pro), ряд иных приложений доступен для загрузки в сети Интернет, например, AtteStat.

Система обработки текстов включает: электронные словари (Викисловарь, Multitrans, ForceMem, ABBYY Lingvo), орфографические (спеллчекеры: MSWord, ispell, aspell, myspell), поисковые системы (Google, Promt, Socrat), системы машинного перевода [3], системы автоматизированного дистрибутивно-статистического анализа (AlchemyAPI, Expert System S.p.A., General Architecture for Text Engineering (GATE), Modular Audio Recognition Framework, Monty Lingua, Natural Language Toolkit (NLTK)) [4].

Современные системы анализа текстов имеют свои особенности, среди них: предпочтение скорости обработки текстов точности семантического и морфологического анализа, выявление смысла текста, реферирование, автоматическое индексирование, эффективная навигация по текстовой базе, статистический частотный анализ словоупотреблений, автоматическая классификация и кластеризация текстов, смысловой поиск и расчет релевантности текстов поисковому запросу.

Одной из процедур обработки текста на естественном языке [4] считается резюмирование. Компрессию и сжатие

текста помогают осуществить процедуры перефразирования (машинный перевод информации в рамках одного и того же языка).

Также можно перечислить и другие программы, которые основаны на обработке текстового материала методом статистического анализа: LightReader, Readability analysis, WordStat, Лингвоанализатор, Система Пропись 4.0, Word-Tabulator v2.2.3, Гамлет, StarLing, TextusPRO, ConcordeRecode, Ventlie Vocable Rhymes, WordStat, Склонятель, Эмпирика, Textarc и др. [4, 5].

Обработка текстовой информации представляет собой платформу, на которой могут быть реализованы не только методы статистической обработки, но и машинного обучения, линейного программирования и т.д., что позволяет решать абсолютно любые задачи управления текстом.

ЛИТЕРАТУРА

1. Системы автоматической обработки текстов [Электронный ресурс]. – Режим доступа: <http://www.osp.ru/os/-2014/01/13039687>, свободный.

2. Бесплатные программы для статистического анализа данных [Электронный ресурс]. – Режим доступа: <http://boris.bilbov.ru/2013/12/01/besplatnyie-programmy-dlya-statisticheskogo-analiza-dannyih>, свободный.

3. Гордукалова, Г. Ф. Анализ информации технологии, методы, организации / Г. Ф. Гордукалова. – СПб : Профессия, 2009. – 508 с.

4. Компьютерные технологии в филологии [Электронный ресурс]. – Режим доступа: <http://ru.wikibooks.org/wiki>, свободный.

5. Программы анализа и лингвистической обработки текстов [Электронный ресурс]. – Режим доступа: <http://www.rvb.ru/soft/catalogue/c01.html>, свободный.