УДК 004.048:004.912

ЛИНГВИСТИЧЕСКИЕ ПРОЦЕССОРЫ И ОБРАБОТКА ТЕКСТОВ НА ЕСТЕСТВЕННЫХ ЯЗЫКАХ

Мушкова В.В.

Пензенский государственный технологический университет, Пенза, e-mail: vikavishenka1@rambler.ru

В статье был проведен анализ работы лингвистических процессоров при обработке текстов на естественном языке, обозначена важность повышения эффективности алгоритмов обработки неструктурированной информации, объём которой увеличивается с каждым годом. Определены задачи, для решения которых служат современные лингвистические процессоры. Были подробно рассмотрены этапы процесса анализа и синтеза текстов на естественном языке при помощи морфологического, синтаксического и семантического анализаторов. Определены задачи и функции анализаторов на каждом из этапов. Рассмотрены наиболее частые способы обработки текста морфологического, синтаксического и семантического анализаторов. Предложены алгоритмы анализа и синтеза текстов на естественном языке. Выявлены факторы, определяющие качество лингвистического процессора. Обозначена актуальность разработки и совершенствования лингвистических процессоров при работе над созданием искусственного интеллекта.

Ключевые слова: лингвистический процессор, естественный язык, искусственный интеллект

LINGUISTIC PROCESSORS AND PROCESSING OF TEXTS ON NATURAL LANGUAGES

Mushkova V.V.

Penza State Technological University, Penza, e-mail: vikavishenka1@rambler.ru

The article is dedicated to the analysis of linguistic processors' work while processing texts in natural language, and the importance of increasing the efficiency of algorithms for processing unstructured information, which volume is growing every year is pointed out as well. The tasks that are to be solved by these modern linguistic processors have been defined. The stages of the process of analyzing and synthesizing texts in natural language with the use of morphological, syntactic and semantic analyzers have been examined properly. The tasks and functions of the analyzers at each stage have been determined. The most frequent ways of processing the text by the use of morphological, syntactic and semantic analyzers have been considered. Algorithms for the analysis and synthesis of texts in natural language have been proposed. The factors that determine the quality of the linguistic processor have been revealed. The necessity of developing and improving linguistic processors while working on the creation of artificial intelligence have been indicated.

Keywords: linguistic processor, natural language, artificial intelligence

Исследования показывают, что совокупный объем данных в 2016 году составил 16 Збайт, а к 2025–му эта цифра увеличится до 163 Збайт. На более чем 80% такая информация является неструктурированной, то есть представляет собой тексты на естественном языке (ЕЯ). Человеку становится с каждым годом труднее ориентироваться в таких огромных потоках поступающей информации.

Многие пользователи имеют определенные служебные обязанности, при выполнении которых не обойтись без постоянного поиска информации в сети Интернет. Обычно необходима вполне конкретная информация, которая называется информационными объектами. Эти объекты различны по своим типам, которые интересуют определённые категории пользователей. Поиск этих объектов в потоке текстов на ЕЯ во многих областях отнимает много времени, что тормозит решение конкретных прикладных задач. Может возникнуть и необходимость в переводе текстов с одного ЕЯ на другой, в конвертации/деконвертации различных языков, создании компьютерных учебников, размеченных корпусов текстов, программ анализа, синтеза, аннотирования и реферирования текстов.

Ключевым элементом информационных систем, ориентированных на ЕЯ, является лингвистический процессор. Он облегчает обмен информацией между экспертной системой и пользователем-человеком.

Лингвистическим процессором называется формальная лингвистическая модель, реализованная на ЭВМ, которая способна понимать и производить тексты на ЕЯ. Основной задачей для лингвистического процессора является извлечение требуемого смысла из предложения или же текста и его интерпретация при помощи набора формальных средств, понятных машине. Система включает в себя три уровня преобразования информации: морфологический, синтаксический и семантический,

каждый из них определяется набором правил. Эти компоненты обеспечивают преобразование текста в набор морфологических, синтаксических и семантических структур и обратно. Обработка текста при помощи лингвистического процессора всегда происходит от простого, легко формализуемого уровня к сложному. В первую очередь осуществляется морфологический анализ, то есть анализ строения слов с подробным описанием их морфологических характеристик. Затем следует синтаксический анализ, который формально описывает порядок слов в предложении и синтаксические связи между ними, указывая слова с описанием частей речи и их набора грамматических характеристик, особое внимание уделяя отношениям синтаксического подчинения. Зачастую синтаксический анализ заканчивается построением синтаксического древа каждого анализируемого предложения. Третьим и самым сложным этапом в процессе анализа информации является построение семантической структуры предложения или текста. Чаще всего она представляется так же древом зависимостей, в котором указываются предметные имена с дугами семантического подчинения (атрибутивные, конъюнкция, дизъюнкция, равенство, неравенство и т.д.).

Для морфологического анализа и морфологического синтеза служит один из компонентов лингвистического процессора – морфологический анализатор. Он позволяет преобразовать морфологические структуры исходной информации в последовательность входящих в текст словоформ с указанием их грамматических характеристик, в том наборе, который характерен для данной части речи. Морфологический синтез же выполняет обратный процесс, то есть преобразует лексемы с указанием их грамматических характеристик в выходной текст, определяя все возможные словоформы по заданной морфологической структуре.

Промежуточным звеном между морфологическим и семантическим анализаторами является синтаксический анализатор. С одной стороны, он отвечает за переход морфологических структур ряда исследуемых словоформ в синтаксическую структуру, с другой – является способом дальнейшего смыслового представления текста. Синтаксический анализатор – программа, осуществляющая сопоставление линейной последовательности лексем с их грамматическими характеристиками. Важнейшей задачей любого синтаксического анализатора является построение синтаксической структуры входного предложения на естественном языке, используя при этом морфологическую информацию о лексемах, полученную при морфологическом анализе, и сведений из словаря. Так же, как и морфологический, синтаксический анализатор включает в себя синтаксический анализ и синтаксический синтез. Синтаксический анализ представляет собой процедуру преобразования морфологической структуры входного предложения на естественном языке в синтаксическую структуру или синтаксическое древо, а синтаксический синтез – это обратное преобразование синтаксической структуры в морфологическую.

Для одновременного морфологического и синтаксического анализа текстов на ЕЯ могут использоваться парсеры, которые позволяют разбить текст на слова и предложения, определить части речи, падеж, род, числа и других грамматические признаки, учитывая контекст слова, привести слово к его начальной форме, определить синтаксические связи слов в предложении, найти подлежащее, сказуемое и т.д. Кроме того, можно включить режим опечаток, чтобы морфологический анализатор исправлял допущенные в тексте орфографические ошибки [1–3].

Семантический анализатор – это программа переработки синтаксической структуры запроса в семантическую при помощи концептуальных графов. Концептуальным графом является способ семантической или, иначе, понятийной репрезентации ситуаций и знаний в моделях понимания естественного языка. Узлами графа становятся те лексические единицы, которые выражают категории и понятия, и соединяются между собой формальными и ассоциативными связями. Ориентация связей в графе всегда направлена сверху вниз – от концептов более высокого уровня обобщения к тем концептам, которые их характеризуют. Элементарный смысл определяется как пара соединенных соседних узлов графа. Подобные связи не всегда именуются, иногда они только фиксируют факт некоторого взаимодействия двух слов (студент-институт, берёза – дерево). Граф состоит из множества связанных между собой элементарных смыслов, которые вступают в дозволенные им комбинации, выявляя лексически активные и пассивные валентности того или иного слова. Связная часть графа, соединяющая два узла, образует подграф. Процесс создания смысловой

категории начинается с того, что из ядерной конструкции предложения выделяется так называемое «ключевое слово». Семантический анализатор является итоговым этапом процесса обработки материала. По завершению работы лингвистического процессора проанализированная информация передается на этап индексирования.

Таким образом, можно составить следующие два алгоритма обработки текста на ЕЯ: алгоритм анализа текста (рис. 1) и алгоритм синтеза текста (рис. 2).

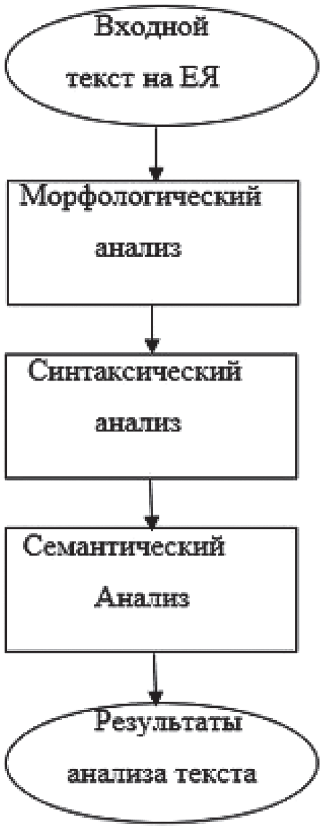

Рис. 1. Алгоритм анализа

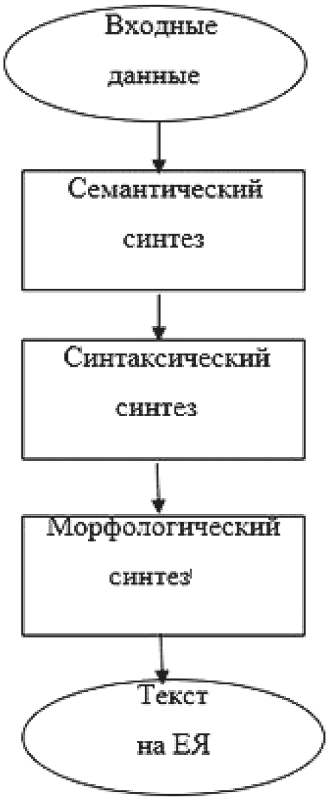

Рис. 2. Алгоритм синтеза

Как видно, синтез текста на естественном языке является обратным алгоритмом анализа текста на естественном языке.

Качество лингвистического процессора в настоящее время определяется рядом важных факторов. А именно, возможностью выделения информационных объектов и связей между ними, избирательностью правил и процедур идентификации, возможностью и трудоемкостью настройки на нужный корпус текстов естественного языка, а также настройки на новые информационные объекты, скоростью, с которой работает лингвистический процессор, временем, которое потребуется для анализа текста.

Разработка лингвистических процессоров является ещё одним шагом к созданию искусственного интеллекта, ведь создание идеального алгоритма точного анализа текста на естественном, человеческом языке станет ключом к пониманию того, как устроен наш язык, а значит и наше мышление.

Список литературы

1. Волкова И.А., Головин И.Г. Лингвистический процессор русского языка: анализ устойчивых словосочетаний // Научные труды SWorld. – 2015. – Т. 2, № 4 (41). – С. 36–46.

2. Ермаков А.Е. Автоматическое извлечение фактов из текстов досье. Опыт установления анафорических связей // Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2007» – Бекасово, 2007, – С. 171–177.

3. Золотарев О.В., Козеренко Е.Б., Шарнин М.М. Принципы построения моделей бизнес-процессов предметной области на основе обработки текстов естественного языка // Вестник РосНОУ. – 2014. – № 4. – С. 82–88.