ЛИТЕРАТУРА
1. Gruber T. R., A Translation Approach to Portable Ontology Specifications. // Knowledge Acquisition. – 1993, 5(2). – pp. 199–220.
2. Муромцев Д.И. Онтологический инжиниринг знаний в системе Protégé // СПб.: СПб ГУ ИТМО, 2007.
3. Антонов И.В., Воронов М.В. Формирование онтологических моделей предметной области для электронных обучающих систем // Информационные технологии в обеспечении нового качества высшего образования. Сборник научных статей. – Кн. 2. – М.: Исследовательский центр проблем качества подготовки специалистов, 2010. – С. 48–55.

Ю.В. БРУТТАН

РОЛЬ ФОРМАЛЬНЫХ СИСТЕМ СЕМАНТИЧЕСКОГО ПРЕДСТАВЛЕНИЯ ТЕКСТОВЫХ ОПИСАНИЙ ПРИ ПРОЕКТИРОВАНИИ ЛИНГВИСТИЧЕСКИХ ПРОЦЕССОРОВ

В статье рассмотрены вопросы понимания компьютером текстовых описаний на естественном языке. Определена роль формальных систем семантического представления текстовых описаний при проектировании лингвистических процессоров.

Пытаясь формализовать понимание компьютером даже довольно простых текстовых описаний, исследователи пришли к выводу о том, что для решения их частных задач необходимо предварительно иметь теоретические методы, относящиеся к произвольным текстам группы естественных языков (ЕЯ), например, русского или английского. Следовательно, требуется разработать такие формальные языки для представления знания о предметных областях и построения семантического представления (СП) принадлежащих им текстов на ЕЯ, чтобы можно было конструировать СП в виде структур, отражающих смысловое содержание этих текстов. Другими словами, нужны формальные языки (или формальные системы, поскольку множество их правильно построенных выражений образует язык) для описания смыслов ЕЯ-текстов, обладающие выразительными возможностями, близкими к возможностям естественного языка.

Формальные модели языка рассматриваются как компоненты различных прикладных систем. Компонента системы, реализующая формальную лингвистическую модель и осуществляющая смысловую обработку текстов на естественном языке, называется лингвистическим процессором (ЛП). Со стороны своего внутреннего устройства лингвистический процессор представляет собой многоуровневый преобразователь, состоящий из трёх уровней пофазного представления текста – морфологического, синтаксического, семантического [1]. Каждый из уровней обслуживается соответствующим компонентом модели – массивом правил и словарями. На каждом из уровней предложение имеет формальный образ, именуемой в дальнейшем структурой – морфологической, синтаксической и семантической структурами.

Под морфологической структурой понимается последовательность входящих в анализируемое предложение слов с указанием части речи и морфологических характеристик (падежа, числа, рода, одушевлённости и т.п.).

Под синтаксической структурой понимается дерево зависимостей, в узлах которого стоят слова данного ЕЯ с указанием части речи и грамматических характеристик, а дуги соответствуют специфичным для данного языка отношениям синтаксического подчинения (например, у синтаксического анализатора Link Parserfor Russian: Xp - связь между началом и концом предложения, Sp - связь между существительным и глаголом и т.д.)

Автором статьи предложена семантическая структура, рассмотренная в [2]. Эту структуру можно назвать «семантическим образом» текстового описания, т.к. она

одновременно является графической моделью текстового описания и содержит в себе семантику исходного текста (на уровне его предикатного представления).

Лингвистический процессор в целом должен обеспечивать выполнение следующих преобразований:

предложение на естественном языке ⇒ морфологическая структура ⇒ синтаксическая структура ⇒ семантическая структура

Реализация лингвистического процессора требует разработки формальных языков (ФЯ) для записи образов предложений на морфологическом, синтаксическом и семантическом уровнях представления; формального понятия структуры предложения для каждой из этих уровней; массивов правил для преобразования структур смежных уровней друг в друга; морфологического, синтаксического и семантического словарей, с включением в них всей информации о каждой лексеме, необходимой для реализации соответствующего преобразования.

Стадии морфологического и синтаксического анализа являются в настоящее время наиболее проработанными лингвистическими этапами процесса обработки ЕЯ-текста. За последние два десятилетия создано, по крайней мере, несколько десятков алгоритмов для разных языков, в том числе 10-12 для русского [3]. Поэтому, остановимся более подробно на наиболее важном и наименее проработанном этапе преобразования текстового описания – этапе формирования семантического представления текстового описания или семантической структуры. Рассмотрим формальный аппарат для отображения смысловой структуры ЕЯ-текстов, относящихся к выбранным предметным областям. Семантическая структура формируется в 2 этапа.

На первом этапе происходит преобразование синтаксической структуры исходного описания на естественном языке в предикатную форму. Язык предикатов, предлагаемый в данной работе складывается из дескрипторного словаря и определённой структуры описания, представляющей собой произвольный набор элементарных высказываний (простых синтагм, предикатных форм) стандартного вида ARB, где A и B – коды терминов вместе с указателями связи, а R – код бинарного отношения отражающего взаимоотношения объектов или признаков рассматриваемой предметной области. Тогда описание любого лингвистического объекта на языке предикатов, принадлежащего некоторой предметной области, будет представлять собой совокупность синтагм вида ARB. Полученный набор синтагм расширяется в результате выполнения формально-логических преобразований над исходной совокупностью.

На втором этапе происходит преобразование сформированной совокупности синтагм в формализованное представление текстового описания в многомерном модельном пространстве, иначе говоря, происходит построение семантического образа текстового описания. Семантический образ текстового описания представляет собой область цветных точек в многомерном модельном пространстве. По осям N-мерного пространства расположены термины тезауруса предметной области. А на пересечении осей располагаются обозначенные цветными точками отношения между соответствующими терминами на осях многомерного пространства. Семантический образ исходного текста на естественном языке хранится в виде матрицы признаков. В качестве признаков семантического образа выбраны координаты точек семантического образа с их кодами цвета. Формула 1 содержит пример семантического образа текста в виде матрицы признаков.

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1N} & RGB_1 \\ x_{21} & x_{22} & \cdots & x_{2N} & RGB_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{M_0 1} & x_{M_0 2} & \cdots & x_{M_0 N} & RGB_{M_0} \end{pmatrix}, \quad (1)$$

где x_{ij} – координаты точек семантического образа в N-мерном модельном пространстве (i=1, 2, …, M_0; j=1, 2, …, N),

RGB_i – цвета точек семантического образа, заданных в виде цветовой модели RGB (i =1, 2, …, M0)

N – количество измерений модельного пространства,

M_0 – количество точек семантического образа текстового описания.

Основное достоинство рассмотренной семантической структуры – единообразное отображение текстовых описаний, имеющих одинаковый смысл, но различающихся конкретными словами или последовательностью предложений в тексте.

На основе рассмотренной семантической структуры текста можно определить технологию формирования семантического представления текстового описания лингвистическим процессором, которая представлена на рисунке 1. Применение данной технологии в различных предметных областях возможно в случае предварительной разработки для каждой из них словарей дескрипторов и предикатов.

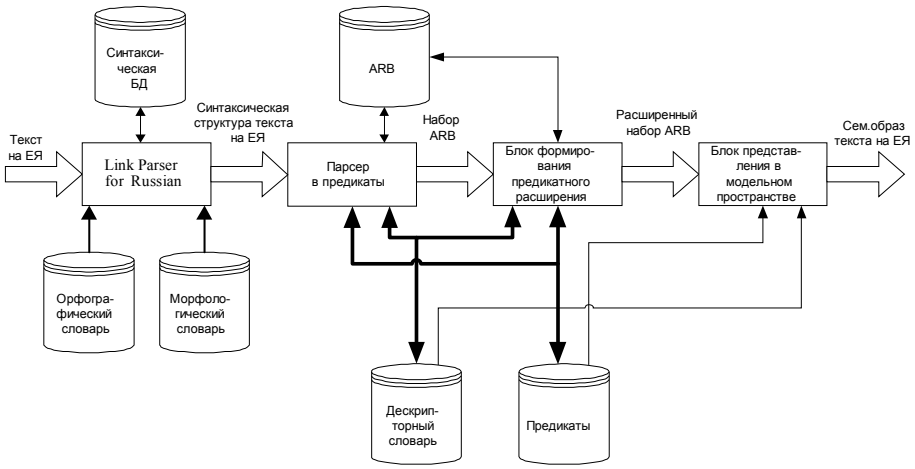

Рис. 1. Технология формирования семантического представления текстового описания ЛП

Таким образом, важнейшую роль при проектировании лингвистического процессора играет, именно, построение формальной модели семантического представления текстовых описаний, которая обеспечивает решение основной задачи ЛП – понимание текста на естественном языке. Предложенную автором модель можно использовать при проектировании прикладных систем с элементами интеллекта. Кроме того, т.к. текстовое описание представлено в виде пространственного изображения, то для него можно применять существующие методы цифровой обработки изображений, что позволит реализовать выполнение операций сравнения и поиска текстовых описаний на новом качественном уровне.

ЛИТЕРАТУРА

1. Апресян Ю.Д. Формальная модель языка и представление лексикографических знаний // Вопросы языкознания. – №6, 1990. – С. 123–139.

2. Бруттан Ю.В. Интеллектуализация поведения компьютеров на основе применения клеточного автомата нового вида // Научно-технические ведомости СПбГПУ. – СПб.: Политехнический университет. – №2, 2007. – С. 225–229.

3. Одинцев Н.В. Обобщенные модели управления. Синтаксический анализатор на основе обобщенных моделей управления // Компьютерная лингвистика и интеллектуальные технологии: Труды международного семинара Диалог'2002. – Т. 2. – М., 2002. – С. 401–406.