УДК 681.3.06

Б.О. Кочконбаева, Н.Р. Абдраева
Ст.преп.ОшТУ,преп.ОшТУ
B.O. Kochkonbaeva, N.R. Abdraeva
Senior teacher OshTU, teacher OshTU

ЛЕКСИЧЕСКИЙ И СИНТАКСИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ И ЕГО ПРИМЕНЕНИЕ В КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЯХ

В данной статье рассматривается вопрос обработки естественного языка. Для этого основными этапами являются лексический, синтаксический и семантический анализ текста. Результатом работы является полученный на языке программирования Delphi лексический и синтаксический анализатор естественного языка.

Ключевые слова: тест, компьютер, лексический, семантический и синтаксический анализ.

THE LEXICAL AND SYNTACTIC ANALYSIS OF TEXTS AND ITS APPLICATION IN COMPUTER TECHNOLOGY

This article addresses the issue of natural language processing. To this end, the main stages are lexical, syntactic and semantic analysis of the text. The work is produced in Delphi programming language and lexical parser of natural language.

Keywords: test, computer, lexical, semantic and syntactic analysis.

Компьютерное моделирование составляет основу современных методов исследования естественного языка. Два важнейших компонента феномена владения языком – это говорение, то есть способность производить тексты, выражающие заданный смысл, и понимание, то есть способность извлекать из текстов смысл, который в них заложен. Модель «Смысл – Текст» есть логическое устройство, имитирующее эти две операции в их простейших проявлениях, связанных исключительно со знанием языка (словаря и грамматики). Некоторые определения. В данной, быстроразвивающейся области, используются некоторые термины, заимствованные из смежных областей знаний. Поэтому во избежание недоразумений здесь целесообразно привести некоторые определения.

Основным из основных направлений искусственного интеллекта является разработка методов, обеспечивающих реализацию процесса общения с ЭВМ на естественном языке. Данное направление получило название «обработка естественного языка» (Natural language processing). Общение с ЭВМ на естественном языке – задача, решение которой предполагает реализацию следующих основных функций: ведение диалога, понимание высказываний, обработка высказываний и генерация выходных высказываний.

Естественный язык (ЕЯ) — язык, словарь и грамматические правила которого обусловлены, практикой применения и не всегда формально зафиксированы.

Онтология является системой понятий, предположительно существующих в некоторой области знаний для обозначения, которых использован определенный естественный язык. В простейшем случае онтология описывает иерархию связных представлений, которые используются для обозначения типовых множеств объектов, обладающих общими признаками.

Обеспечение взаимодействия с ЭВМ на естественном языке является важнейшей задачей исследований по искусственному интеллекту. Сейчас обработка языка активно включаются в различные сферы нашей жизни, способствуя ускорению процессов информационного обмена в различных предметных областях, что привело к развитию

проблемно-ориентированных систем понимания текста. При этом наиболее остро проявилась проблема разрешения языковой неоднозначности, а также проблема учета информации об иерархии понятий и терминов определенной предметной области. Первая проблема обусловлена многозначностью слов естественного языка, ошибками распознавания отдельных слов и синтаксическими неточностями в тексте. Вторая ведет к терминологической путанице, возникающей из-за разницы в толковании терминов у системы и пользователя. Решение этих проблем связано с адекватным отображением естественного языка во внутреннее машинное представление. Для этого следует эффективно использовать всю доступную априорную информацию, включая синтаксис, семантику и прагматику.

Как правило, подходы к представлению и обработке естественного языка используют только два вида информации: синтаксическую и семантическую. Причем: основной упор делается на синтаксис, т.е. методы грамматического разбора. Синтаксический анализ становится самоцелью и приводит к построению грамматически правильных предложений, которые, однако, могут содержать, смысловую неоднозначность. В результате многолетних исследований в области обработки естественного языка и речи было установлено, что для решения проблемы, неоднозначности необходимо использовать информацию о соотнесении знаков естественного языка, объектов и событий реальной действительности, к которым относятся – семантическая и прагматическая информация, и которые представляют собой по существу информацию о предметной области. Стало очевидным, что сложность понимания и методы обработки естественного языка определяются не только структурой и особенностями входного текста, но и представлением о предметной области, в рамках которой осуществляется человеко-машинное взаимодействие.

Существует достаточно обширный набор средств представления знаний о предметной области, наиболее эффективным на сегодняшний день считается онтология. Применение этих средств для представления семантической и прагматической информации в области обработки естественного языка является актуальной темой исследования, поскольку ведет к разрешению проблем языковой неоднозначности и учета иерархии понятий предметной области при обработке текста.

На основе вышесказанного основной целью является разработка методов разрешения, неоднозначности естественного языка и учета иерархии понятий при представлении и обработке естественного языка. Для достижения поставленной цели решены следующие задачи:

1. Анализ основных подходов к представлению и обработке естественного языка;
2. Построение эффективной модели представления, и обработки естественного языка;
3. Разработка методов эффективного семантико-прагматического анализа.

Для решения поставленных задач используются методы теории информации, теории множеств, экспертного, статистического и эвристического анализа, а также методы итерационного поиска. Компьютерная реализация - разработанных алгоритмов производилась на объектно-ориентированном языке программирования Delphi. Эта программа в данный момент выполняет лексический и семантический анализ текста. Базовым естественным языком для исследования был выбран кыргызский язык. На фазе лексического анализа входная программа, представляющая собой поток литер, разбивается на лексемы - слова в соответствии с определениями языка. Основными формализмами, лежащим в основе реализации лексических анализаторов, являются конечные автоматы и регулярные выражения. Лексический анализатор может работать в двух основных режимах: либо как подпрограмма, вызываемая синтаксическим анализатором для получения очередной лексемы, либо как полный проход, результатом которого является файл лексем.

В процессе выделения лексем лексический анализатор может, как самостоятельно строить таблицы объектов (идентификаторов, строк, чисел и т.д.), так и выдавать значения для каждой лексемы при очередном к нему обращении. В этом случае таблицы объектов строятся в последующих фазах (например, в процессе синтаксического анализа).

На этапе лексического анализа обнаруживаются некоторые ошибки (недопустимые символы, неправильная запись чисел, идентификаторов и др.).

Основная задача синтаксического анализа - разбор структуры программы. Как правило, под структурой понимается дерево, соответствующее разбору в контекстно-свободной грамматике языка. В настоящее время чаще всего используется либо LL(1)-анализ (и его вариант - рекурсивный спуск), либо LR(1)-анализ и его варианты (LR(0), SLR(1), LALR(1) и другие). Рекурсивный спуск чаще используется при ручном программировании синтаксического анализатора, LR(1) - при использовании систем автоматического построения синтаксических анализаторов.

Результатом синтаксического анализа является синтаксическое дерево со ссылками на таблицы объектов. В процессе синтаксического анализа также обнаруживаются ошибки, связанные со структурой программы.

На этапе контекстного анализа выявляются зависимости между частями программы, которые не могут быть описаны контекстно-свободным синтаксисом. Это в основном связи «описание-использование», в частности, анализ типов объектов, анализ областей видимости, соответствие параметров, метки и другие. В процессе контекстного анализа таблицы объектов пополняются информацией об описаниях (свойствах) объектов.

Основным формализмом, использующимся при контекстном анализе, является аппарат атрибутных грамматик. Результатом контекстного анализа является атрибутированное дерево программы. Информация об объектах может быть, как рассредоточена в самом дереве, так и сосредоточена в отдельных таблицах объектов. В процессе контекстного анализа также могут быть обнаружены ошибки, связанные с неправильным использованием объектов.

Затем программа может быть переведена во внутреннее представление. Это делается для целей оптимизации и/или удобства генерации кода. Еще одной целью преобразования программы во внутреннее представление является желание иметь переносимый компилятор. Тогда только последняя фаза (генерация кода) является машинно-зависимой. В качестве внутреннего представления может использоваться префиксная или постфиксная запись, ориентированный граф, тройки, четверки и другие.

Сейчас значительная часть содержания сети предназначена для чтения человеком, а не для осмысленного манипулирования этим содержанием с помощью компьютерных программ. Начиная с 2001 года, во многих странах мира ведется разработка семантической паутины (Semanticweb), представляющей собой надстройку над существующей всемирной паутиной. Семантическая паутина призвана сделать размещённую в ней информацию более понятной для компьютеров. Для этого необходимо создать специальные формализмы для записи семантической информации, технологию работы с ними и разработать обширные текстовые ресурсы, записанные в этих формализмах. В этом процессе один из ключевых моментов состоит в семантическойинтерпретации текстов, написанных на естественном языке. Эта семантическая интерпретация может осуществляться на базе естественно-языковых систем и последующим этапом нашей программы.

В заключении хочу сказать, что:

- компьютерные технологиизавтрашнего дня нуждаются в фундаментальных исследованияхв области моделирования естественного языка.
- модели, которые существуют сегодня, могут служить хорошей основой для этих исследований.
- уже сегодня существуют полезные приложения, основанные на этих моделях.

Литература:

1. Апресян Ю.Д., Богуславский ИМ., Иомдин JI.JI. и др. Лингвистический процессор для сложных информационных систем: М.: Наука, 1992
2. Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. и др. Лингвистическое обеспечение системы Этап-2, М.: Наука, 1989. — 295с.

3. Бехтель Э. Е., Бехтель А. Э. Контекстуальное опознание. – СПб.: Питер, 2005. – 336 с.
4. Михайлов А. Урок нечеловеческого языка. Роботы учатся говорить со своими хозяевами // РБК daily. – N. 124. – 16 июля 2010.
5. Мельников Г.П. Системная типология языков. - М.: Наука, 2003. – 532 с.
6. Рыков В.В. Обработка нечисловой информации. Управление знаниями. – М.: МФТИ, 2008.