В. Князев

ИСПОЛЬЗОВАНИЕ СИНТАКСИЧЕСКИХ ДЕРЕВЬЕВ ДЛЯ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Синтаксический анализ – одна из задач обработки естественного языка, которая заключается в разделении предложения или иной последовательности слов на сегменты и объединении их в какую-либо структуру данных (как правило дерево), которая показывает синтаксическую связь ее элементов друг с другом. Синтаксический анализ может быть применен для любого языка, описанного формальной грамматикой.

В данном направлении основополагающей концепцией является предложенная профессором Массачусетского технологического института лингвистом Ноамом Хомским классификация формальных языков и грамматик, называемая Иерархией Хомского. В ней выделяются 4 уровня, от наиболее общей грамматики до самой строгой, при этом грамматика более высокого уровня, по определению, относится ко всем грамматикам более низкого уровня. Формальная грамматика представляет из себя набор символов и правил, которые показывают, как можно преобразовать один набор символов в другой.

Символы бывают терминальными и нетерминальными в зависимости от того, должны ли они быть заменены. В конечном счете, при применении всех правил должны остаться лишь терминальные символы. В случае с естественными языками в качестве терминальных символов выступают слова, а нетерминальных – различные синтаксические единицы, такие как глагольная или именная группа. Чем более строгая грамматика, тем легче создать алгоритм для синтаксического анализа, но при этом более общие правила позволяют определить более сложные аспекты языков. Чаще всего при описании естественных языков, как и многих языков программирования, используется контекстно-свободная грамматика 2 типа. Однако стоит отметить, что эти языки являются слишком сложными для того чтобы в полной мере вписаться в рамки идеализированной модели формальной грамматики, и для их обработки требуются дополнительные шаги, например, семантический анализ. Это связано, в первую очередь, с тем, что значения некоторых слов можно понять исключительно по контексту: многозначные слова, омонимы, слова, употребленные в переносном значении.

Синтаксический анализ происходит следующим образом: синтаксический анализатор, или же *парсер*, обычно сочетается с лексическим анализатором, который производит токенизацию – процесс преобразования

последовательности символов в последовательность токенов, имеющих определенное значение в языке. Затем эти токены используются для построения синтаксического дерева по заданным правилам. Заключительный этап – вычисление. Как правило, оно производится слева направо, снизу вверх, в таком случае синтаксический анализатор называется *нисходящим LL-анализатором*.