УДК 519.682

Д. В. Личаргин

ПОРОЖДЕНИЕ ДЕРЕВА СОСТОЯНИЙ НА ОСНОВЕ ПОРОЖДАЮЩИХ ГРАММАТИК НАД ДЕРЕВЬЯМИ СТРОК

Рассмотрен принцип порождения дерева состояний на основе порождающих грамматик над деревьями строк над такими объектами, как предложения естественного языка, а также двумерные и трехмерные образы. Рассматривается представление объекта как леса, включающего деревья разных срезов этого объекта в целях моделирования сложных систем.

Ключевые слова: порождение естественного языка, порождающие грамматики, семантика.

Проблема порождения предложений естественного языка является одной из важных проблем семантики и информатики [1–7]. Проблема порождения дерева состояний рассматривается в информатике и системном анализе весьма широко. Относительно вопроса генерации дерева осмысленных фраз эта проблема связывается в первую очередь с методом генерации предложений при помощи порождающих грамматик Хомского. Порождающие грамматики успешно применяются в таких программах, как системы перевода, экспертные системы, системы проверки орфографии и т. п.

Основной идеей данной статьи является анализ перспективы использования порождающих грамматик не над строками, а над деревьями строк. В связи с этим возможно более эффективное решение, с одной стороны, задач порождения грамматически и семантически осмысленной речи, а с другой – повышения эффективности различных аспектов анализа и синтеза образов.

Актуальность проблемы эффективной генерации осмысленных конструкций языка и двумерных и трехмерных образов является общепризнанной и связана с потребностями лингвистического и иного программного обеспечения.

Цель работы состоит в обосновании необходимости применения порождающих грамматик над деревьями как средства генерации осмысленной речи с учетом более разнородного контекста. Новизна работы состоит в применении порождающих грамматик не над строками, а над деревьями строк.

Как известно, стандартные порождающие грамматики над строками имеют вид четверки: G <S, T, N, R>, где S – начальный символ порождающей грамматики; T – множество терминальных символов; N – множество нетерминальных символов; R – множество правил трансформации одной строки в другую.

Для порождающих грамматик над деревьями строки символов t и n заменяются деревьями (или лесом – деревьями с тождественными узлами). t = t <t’, t”, …, t^n>, где t’ = t’ <t^1, t^2, …, t^m> и т. д., n = n <n’, n”, …, n^n>, где n’ = n’ <n^1, n^2, …, n^m> и т. д.

Одной из основных особенностей любой системы является иерархия элементов системы. При этом иерархические отношения иногда могут составлять множество иерархий различных срезов рассмотрения системы. Например, сложение трех систем: высказывания в рамках распространенного повествования, высказывание с целью заказать чай и высказывание с целью поддержания вежливого разговора может дать осмысленные предложения естественного языка. При этом для генерации таких сложных систем с несколькими целями и срезами рассмотрения необходимо использовать более сложные средства, чем порождающие грамматики над строками символов. Предлагается использовать порождающие грамматики над деревьями строк в целях генерации дерева возможных высказываний естественного языка.

Порождающая грамматика над деревьями строк строится следующим образом. Пусть A<…B<…C1→C2…>,…, B’<…C1’→C2’>…> – правило порождающей грамматики над деревьями из множества таких правил с деревьями строк терминальных символов T и нетерминальных символов N; → – символ перехода одной строки в другую; S<> – начальный символ порождающей грамматики над деревьями.

Углубление дерева состояний другого генерируемого дерева или леса строк состоит на каждом этапе в умножении получаемого генерируемого дерева на правило порождающей грамматики.

Можно рассмотреть также деревья разнородной информации A<B{B1, B2}, C{C1, C2}> = {A<B1,C1>, A<B1,C2>,A<B2,C1>,A<B2,C2>} = {A<B1,C1>,A<B1,C2>, A<B2,C{C1, C2}>}. Таким образом, дерево состояний системы может быть вложено в дерево элементов системы и наоборот.

Как результат, высказывание может рассматриваться в виде объединения (сложения) деревьев разных срезов рассмотрения над единым пространством (деревом) точек слов естественного языка [4–6].

Пусть дано дерево A <B <B’<…>, B”<…>,…, B’”<…>>, C <C’<…>, C”<…>, C’”<…>>, …, D <D’<…>, D”<…>, …, D’”<…>>> или коротко A <…B <…B”…>…>, тогда лес деревьев рассмотрим как множество деревьев с тождественными узлами на множестве узлов этих деревьев: F<A<…B<…B”(=L1)…>…>,X<…Y<…Y”(=L1)…>…>,…>, где L1 – тождественный узел первых двух деревьев вышеприведенного примера .

Рассмотрим пример дерева комбинаций шахматной партии: Доска <Колонка [1] <Клетка [1], Клетка [2],…>,…>, такое дерево формируется посредством умножения позиции на доске на множество правил возможных полуходов.

Ход конем будет иметь следующий вид: Доска <…Колонка [X] <…Клетка[Y] <Конь→Пусто>>, …, Колонка [(X + 1)or(X–1)] <Клетка[(Y + 2)or(Y-2)] <Пусто→ Конь>…>…>.

Генерация, например, образа стула предполагает также потенциальный образ человека на этом стуле. Стул <Сидение, Ножки, Спинка, Человек(= L1) <Руки(= L2), Ноги(= L3),

Туловище(=L4), Голова(=L5)>> + Джентльмен(=L1) <Тело <Руки(=L2), Ноги(=L3), Туловище(=L4), Голова(=L5)>, Одежда <Пиджак <Туловище<= L4>>, Ботинки, Цилиндр <Голова(=L5)>>> = Рисунок<Стул<…>, Джентльмен<…>,…>.

Принцип свертки или сложения образов заключается в следующем: семантически схожие элементы – узлы деревьев – объявляются тождественными; в случае наличия нескольких вариантов свертки строится дополнительное подпространство возможных состояний системы – результата сложения деревьев элементов системы и порождения деревьев состояний системы.

Предложение естественного языка также может быть представлено в виде дерева. Например, дерево грамматического разбора предложения упрощенно может иметь следующий вид: Предложение <Вводное слово, Обстоятельство, Субъект <Определитель, Определение <Наречие степени, Группа прилагательного>, Именная часть>, Предикат <Модальность, Обстоятельство, Глагольная часть>, Объект <Определитель, Определение <Наречие степени, Группа прилагательного>, Именная часть>, Обстоятельство>.

Данное дерево может быть прибавлено к (свернуто с) деревом семантического анализа, например, Тема «Здания» <Отношение-Существо-Здание {входить в, строить}, Свойство-Здание {мраморный, многоэтажный}, Здание {дом, библиотека}, Обстоятельство 1 <с/без {с, без}, Сущность-Здание/Комнаты {коридор, зал}>, Обстоятельство 2 <с/без {с, без}, Свойство-Предмет(Сущность-Здание/Архитектурный элемент {большой, красивый}), Сущность-Здание/Архитектурный элемент {стена, угол}>>.

Дерево следующего вида может быть использовано для генерации предложений естественного языка.

1. Субъект – существо (этот …/ человек / мужчина / женщина).

2. Модальность – действие над отношением (хотеть / желать / любить / обожать).

3. Предикат – действие с одеждой (покупать / получать / примерять / носить).

4. Объект – одежда (этот … / джинсы / свитер / футболка).

Данное дерево может быть умножено на следующее правило порождающей грамматики.

1. 0 →Этот.

2. 0 → Атрибут – свойство одежды (стильный / модный / клетчатый).

3. Объект – Одежда (Этот … → 0 / джинсы / свитер / футболка).

В результате получается предложение следующего вида: «этот человек хочет получить этот модный свитер» или «эта женщина желает купить эту клетчатую футболку».

Можно предположить, что проблемы распознавания образов, анализа естественного языка и ряд других могут быть эффективно решены только на основе их совместного синтетического рассмотрения. Так, например, для перевода выражения «up-link communication» как «связь со спутником» необходимо использовать визуальный образ того, о чем говорится в тексте. Таким образом, в системе перевода при переводе текста должен наращиваться семантико-визуальный образ повествования, без которого невозможен перевод, приближенный к переводу человеком.

Для реализации вышеупомянутых принципов предполагается начать разработку словаря семантических деревьев разнородных данных: образов, шаблонов построения предложений, алгоритмов и т. п. В основу системы будет положен уже существующий словарь порождения высказывании в программе «Электронный словарь».

Вывод данной работы состоит в том, что порождающие грамматики над деревьями строк являются эффективным средством порождения деревьев состояний таких систем, как предложение естественного языка и семантически нагруженный образ. Предполагается применение порождающих грамматик над деревьями строк на основе «Словаря семантических деревьев», представляющего собой классификацию разнородных семантических данных.

Библиографические ссылки

1. Агамджанова В. И. Контекстуальная избыточность лексического значения слова. М. : Выс. шк., 1977.

2. Апресян Ю. Д. Идеи и методы современной структурной лингвистики. М. : Наука, 1966.

3. Вердиева З. Н. Семантические поля в современном английском языке. М. : Высш. шк., 1986.

4. Личаргин Д. В. Операции над семами слов естественного языка в машинном переводе // Тр. конф. молодых ученых ; Ин-т вычислит. моделирования СО РАН. Красноярск, 2003. С. 23–31.

5. Личаргин Д. В. Устранение семантического шума как средство адекватного перевода // Вопросы теории и практики перевода : тр. Всерос. конф. Пенза, 2003. С. 90–92.

6. Личаргин Д. В. Порождение фраз естественного языка в рамках задачи построения естественно-языкового интерфейса с программным обеспечением // Проблемы информации региона (ПИР-2003) : материалы восьмой Всерос. конф. Т. 2. Красноярск, 2003. С 152–156.

7. Никитин М. В. Лексическое значение слова. М. : Высш. шк., 1983.

D. V. Lichargin

THE GENERATION OF STATES TREE ON THE BASIS OF THE GENERATIVE GRAMMARS OVER THE TREES OF STRINGS

The principle of states trees generation based on the generative grammars over trees of strings over such objects as the sentences of the natural languages, as well as two and three-dimensional images is considered. The presentation of the object as a forest including the trees of different layouts of the object for the purpose of complex systems modeling is considered.

Keywords: natural language generation, generative grammars, semantics.