25 июня 2008 г.

П. В. Егоров

ОБ ОДНОВРЕМЕННОМ ПРИМЕНЕНИИ НЕСКОЛЬКИХ НЕЗАВИСИМЫХ ПРЕОБРАЗОВАНИЙ ГРАМАТИК

1. Введение

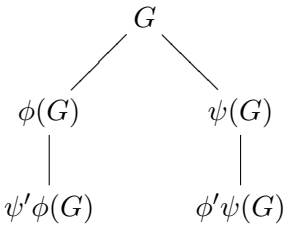
В данной работе представлен очередной шаг к созданию расширяемых синтаксических анализаторов формальных языков – рассмотрены некоторые важные вопросы, касающиеся одновременного применения нескольких независимых преобразований к грамматике.

Ниже будет описана постановка задачи и сформулированы рассматриваемые в данной работе вопросы. Далее будут последовательно введены понятия, на основе которых затем будут даны ответы на поставленные вопросы.

2. Постановка задачи

Пусть у нас есть грамматика G и два преобразования этой конкретной грамматики – ϕ и ψ. Каждое преобразование оперирует правилами вывода – удаляет одни и добавляет другие. Если применить к грамматике одно из преобразований, то встаёт вопрос о том как к полученной грамматике $\phi(G) \neq G$, применять ψ, предназначенное для применения к G. В некоторых случаях это будет не возможно. В некоторых других случаях можно немного изменив ψ получить новое преобразование ψ‘, которое уже предназначено для применения к $\phi(G)$ и в некотором смысле преобразует $\phi(G)$ тем же способом, что и ψ преобразует G.

Это можно изобразить графически следующим образом:



В это работе рассмотрены следующие вопросы. Как нужно изменять исходное преобразование, чтобы его можно было применить после другого преобразования? Какие требования логично предъявлять к этому способу модификации исходного преобразования? Будет ли зависеть результат применения обоих преобразований от порядка их применения?

3. Размеченная грамматика

Пусть есть некоторая грамматика $G = (V, \Sigma, P, S)$ и p – это некоторое правило вывода из P. Введём несколько вспомогательных обозначений.

Через $len(p)$ будем обозначать длину правой части правила вывода p.

Через $p(i)$, при $1 \leq i \leq len(p)$, будем обозначать i-ый символ правой части правила вывода p.

Через $p(0)$ будем обозначать символ левой части правила вывода p.

Определение 3.1. *Разметкой грамматики G будем называть функцию m, которая каждому правилу вывода грамматики G будет ставить в соответствие кортеж, состоящий из нулей и единиц, длиной в количество символов в правой части своего аргумента. Другми словами для любого правила p $m(p)$ есть кортеж*

$$(m_1, m_2, \ldots m_{len(p)}), \quad \textit{где} \quad m_i \in \{0, 1\}$$

Кортеж $m(p)$ будем называть разметкой правила p.

Для удобства введём слудеющее обозначение для элементов кортежа:

$m(p) = (m(p, 1), m(p, 2), \ldots m(p, len(p)))$.

Значение $m(p, i)$ будем называть разметкой i-ого символа правой части правила p.

Определение 3.2. *Определим множество размеченных грамматик Γ_M как множество всевозможных пар (G, m), где $G \in \Gamma$, а m – разметка G. Саму пару (G, m) будем называть размеченной грамматикой.*

Разметку правила вывода будем обозначать с помощью верхнего индекса у символов правой части правила вывода. Например для правила вывода $p : A \to BCde$ и разметки $m(p) = (0, 1, 0, 1)$, размеченное правило вывода будем обозначать так: $p : A \to B^0C^1d^0e^1$

4. Синтаксический анализатор

Определение 4.1. *Пусть L – некоторый язык. Тогда синтаксическим анализатором G будем называть произвольную функцию, определенную на L.*

Ниже будет определён синтаксический анализатор $ct\langle G, m\rangle$, управляемый размеченной грамматикой. Он будет в качестве результата возвращать конструкцию, очень похожую на классическое дерево вывода слова в грамматике G. Разметка будет им использоваться для того, чтобы вырезать некоторые узлы из классического дерева вывода. Рассмотрим всё по порядку.

Каждому узлу кроме корня в классическом дереве вывода можно сопоставить значение разметки правила, в результате применения которого появился данный узел. Такое сопоставление будем называть разметкой узлов дерева вывода. Для единообразия доопределим разметку на корне дерева единицей. Чтобы пояснить идею разметки дерева вывода, обратимся к примеру:

Пусть дана следующая грамматика:

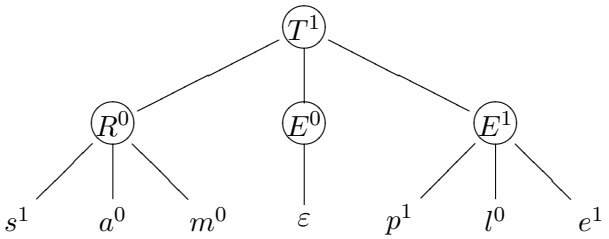
$$T \to R^0 E^0 E^1$$

$$R \to s^1 a^0 m^0$$

$$E \to \varepsilon$$

$$E \to p^1 l^0 e^1$$

Соответствующее дерево вывода слова «sample» в данной грамматике будет выглядеть так:



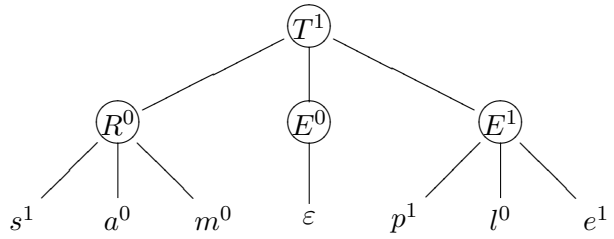
Определение 4.2. *Вспомогательным деревом вывода слова w в грамматике G, назовём дерево, полученное из классического дерева вывода слова w в грамматике G, в результате удаления всех листьев, помеченных ε.*

Определение 4.3. *Сокращённым деревом вывода слова w в грамматике G, с разметкой m назовём дерево, полученное из вспомогательного дерева вывода слова w в грамматике G, путём вырезания из дерева всех узлов, размеченных нулями. Ниже дано определение вырезания узла из дерева.*

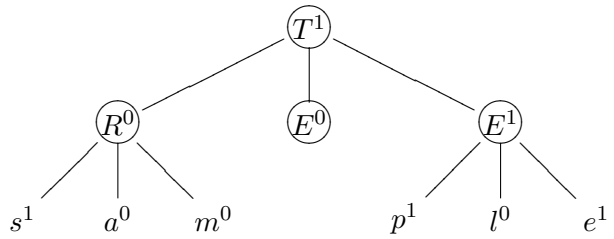
Определение 4.4. *Пусть у узла p_1 есть сыновья $n_1, n_2, \ldots n_k$. А у узла $n_i, (1 \le i \le k)$ есть сыновья $c_1, c_2, \ldots c_m$. Вырезанием узла n_i из дерева, называется операция замены в списке сыновей p_1 узла n_i на список своих сыновей $c_1, c_2, \ldots c_m$.*

Пример 4.1. *Сокращённое дерево вывода для классического дерева вывода из предыдущего примера.*

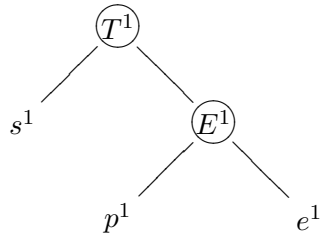
Исходное классическое дерево вывода:



Вспомогательное дерево вывода:



Сокращённое дерево вывода:



Определение 4.5. *Пусть CT – это множество всех сокращённых деревьев вывода. Определим $ct\langle G, m\rangle$ как синтаксический анализатор, который слову ставит в соответствие его сокращённое дерево вывода в грамматике G с разметкой m.*

Далее мы будем изучать свойства данных синтаксических анализаторов, при применении к грамматике расширяющих преобразований.

5. Расширяющие преобразования AE

Определение 5.1. *Расширяющим преобразованием размеченных грамматик назовём функцию $f : \Gamma_M \to \Gamma_M$ такую, что*

1. $L(f(G,m)) \subset L(G)$.
2. $\forall w \in L(G), ct\langle G,m\rangle(w) = ct\langle f(G,m)\rangle$

Другими словами, расширяющее преобразование увеличивает язык грамматики, но не изменяет действия синтаксического анализатора ct на языке исходной грамматики.

Определим класс расширяющих преобразований Add. Произвольное правило вывода с произвольной же своей разметкой задаёт преобразование, добавляющее это правило в размеченную грамматику. Возможно, перед этим понадобится добавить в множество терминалов все ещё отсутствующие там терминалы из правой части правила вывода. Множество таких преобразований обозначим Add.

Область определения преобразования из Add – это множество грамматик, в которых добавляемое правило отсутсвует, а все нетерминалы правила вывода уже определены в грамматике.

Далее определим класс расширяющих преобразований $Extract$. Рассмотрим четвёрку (p,l,r,B), где p – правило вывода $A \to \alpha\beta\gamma$, $l = |\alpha|$, $r = |\alpha|+|\beta|$, а B – произвольный символ, не являющийся терминальным в исходной грамматике. Эта четвёрка определяет преобразование грамматики, удаляющее правило вывода p и добавляющее вместо него два правила вывода: $A \to \alpha B\gamma$ и $B \to \beta$. Причём разметка певого на участках α и γ совпадает с разметкой удаляемого правила на тех же участках, разметка символ B в первом добавляемом правиле – 0, а разметка второго добавляемого правила совпадает с разметкой соответствующего участка удаляемого правила вывода. Множество таких преобразований обозначим $Extract$.

Область определения преобразования из $Extract$ – это множество грамматик, в которых присутствует правило вывода p и отсутствует правило вывода $B \to \beta$, причём терминал B может быть и не определён в исходной грамматике.

Пример 1. *Пример примерения преобразования Extract.*

Исходную грамматику с одним правилом вывода $p : S \to s^0a^1m^0p^1l^0e^1$ преобразование $(p,2,4,B)$ превратит в следующую грамматику:

$$S \to s^0a^1B^0l^0e^1$$

$$B \to m^0p^1$$

Определение 5.2. *Введём два обозначения:*
$AE = Add \cup Extract$
$AE^* = \langle AE\rangle$ *– транзитивное замыкание AE.*

6. Независимые преобразования

Определение 6.1. *Пусть ϕ и ψ – это преобразования из AE. Определим операцию ϕ/ψ следующим образом:*

1. *Если $\phi \in Add$ или $\psi \in Add$, то $\phi/\psi = \phi$*
2. *Если $\phi = (p_1, l_1, r_1, N_1)$, $\psi = (p_2, l_2, r_2, N_2)$, а $p_1 \neq p_2$, то $\phi/\psi = \phi$.*
3. *Иначе, введём дополнительные обозначения: $\phi = (p, l_1, r_1, N_1)$, $\psi = (p, l_2, r_2, N_2)$,*

   t, b – правила вывода, добавляемые преобразованием ψ, причём b – это то правило вывода, левая часть которого N_2.

   $\Delta = l_2 - r_2 + 1$.

   Этот случай разбивается на следующие подслучаи:

   (a) *если $l_1 \leq l_2 \wedge r_1 > r_2 \vee l_1 < l_2 \wedge r_1 \geq r_2$, то $\phi/\psi = (t, l_1, r_1 + \Delta, N_1)$*
   (b) *если $r_1 \leq l_2$, то $\phi/\psi = (t, l_1, r_1, N_1)$*
   (c) *если $l_1 \geq l_2$, то $\phi/\psi = (t, l_1 + \Delta, r_1 + \Delta, N_1)$*
   (d) *если $l_1 \geq l_2 \wedge r_1 < r_2 \vee l_1 > l_2 \wedge r_1 \leq r_2$, то $\phi/\psi = (b, l_1 - l_2, r_1 - l_2, N_1)$*
   (e) *иначе значение не определено.*

Для простоты обозначений будем считать, что операция / имеет меньший приоритет, чем операция композиции функций. То есть $\phi_2\phi_1/\psi_2\psi_1 = (\phi_2\phi_1)/(\psi_2\psi_1)$.

По аналогии с операцией деления будем применять ещё и двухстрочную запись: $\phi/\psi = \frac{\phi}{\psi}$.

Определение 6.2. *Доопределим рекурсивно операцию ϕ/ψ на преобразования из AE^*.*

Если $\exists \phi_1 \in AE, \phi_2 \in AE^* : \phi = \phi_2\phi_1$, то

$$\frac{\phi}{\psi} = \frac{\phi_2}{\psi/\phi_1}\,\frac{\phi_1}{\psi} \tag{1}$$

Иначе $\exists \psi_1 \in AE, \psi_2 \in AE^* : \psi = \psi_2\psi_1$. В этом случае

$$\frac{\phi}{\psi} = \frac{\phi/\psi_1}{\psi_2} \tag{2}$$

В обоих тождествах считаем, что если значение хоть одного из подвыражений не определено, то значение ϕ/ψ также не определено.

Определение 6.3. *Пусть* $\phi = \phi_n\phi_{n-1}\dots\phi_1$, *где* $\phi_i \in AE$. *Тогда* n *будем называть сложностью преобразования* ϕ.

Лемма 1. *Пусть* $\phi \in AE^*$ *– преобразование сложности* n. *Тогда для любого* $\psi \in AE^*$, ϕ/ψ *также имеет сложность* n.

Лемма 6.1. *Для любых* ψ_1, *из* AE *и* ϕ, ψ_2 *из* AE^* *верно следующее равенство:*

$$\frac{\phi}{\psi_2\psi_1} = \frac{\phi/\psi_1}{\psi_2}$$

Доказательство. Доказательство будет проведено индукцией по сложности преобразования ϕ. База индукции тривиально следует из определения операции /.

Предположение индукции. Пусть для ϕ сложности не более n справедливо равенство:

$$\frac{\phi}{\psi_2\psi_1} = \frac{\phi/\psi_1}{\psi_2} \tag{3}$$

Шаг индукции. $\exists\phi_1 \in AE, \phi_2 \in AE^* : \phi = \phi_2\phi_1$.

$$\frac{\phi}{\psi_2\psi_1} = \frac{\phi_2\phi_1}{\psi_2\psi_1} \overset{(1)}{=} \frac{\phi_2}{\psi_2\psi_1/\phi_1}\,\frac{\phi_1}{\psi_2\psi_1} \overset{(1)}{=} \frac{\phi_2}{\frac{\psi_2}{\phi_1/\psi_1}\,\frac{\psi_1}{\phi_1}}\,\frac{\phi_1/\psi_1}{\psi_2} \overset{(3)}{=}$$

$$= \frac{\left(\frac{\phi_2}{\psi_1/\phi_1}\right)}{\left(\frac{\psi_2}{\phi_1/\psi_1}\right)}\left(\frac{\phi_1/\psi_1}{\psi_2}\right) \overset{(1)}{=} \frac{\frac{\phi_2}{\psi_1/\phi_1}\,\frac{\phi_1}{\psi_1}}{\psi_2} \overset{(1)}{=} \frac{\phi_2\phi_1/\psi_1}{\psi_2} = \frac{\phi/\psi_1}{\psi_2}$$

Теорема 6.1. $\frac{\phi}{\psi}\psi = \frac{\psi}{\phi}\phi$

Доказательство. Доказательство будем вести индукцией по сложности преобразований ϕ и psi.

База индукции. Пусть ϕ и ψ из AE, рассмотрим единственный нетривиальный случай – это $\phi = (t, l_1, r_1, N_1), \psi = (t, l_2, r_2, N_2) \in Extract$. Существует два принципиально различных случая. Рассмотрим их оба:

1 случай. $l_1 \le l_2 \wedge r_1 > r_2 \vee l_1 < l_2 \wedge r_1 \ge r_2$. Преобразование ψ удаляет одно правило вывода $t : A \to \alpha_1\alpha_2\beta\gamma_2\gamma_1$ и вместо него добавляет два новых правила: $A \to \alpha_1\alpha_2N_2\gamma_2\gamma_1$, $N_2 \to \beta$.

Преобразование ϕ/ψ удалит в этой грамматике первое правило вывода и заменит его двумя следующими правилами:

$A \to \alpha_1N_1\gamma_1$, $N_1 \to \alpha_2N_2\gamma_2$,

В результате преобразование $(\phi/\psi)\psi$ удалит из исходной грамматики одно правило вывода t и добавит три следующих правила вывода:

$A \to \alpha_1 N_1 \gamma_1$, $N_1 \to \alpha_2 N_2 \gamma_2$, $N_2 \to \beta$.

Аналогичными рассуждениями несложно убедиться, что преобразование $(\psi/\phi)\phi$ делает в точности тоже самое.

2 случай. $r_1 \leq l_2$. Рассуждениями, аналогичными проделанным в доказательстве предыдущего случая, легко показать, что преобразования $(\phi/\psi)\psi$ и $(\psi/\phi)\phi$ совпадают.

Предположение индукции. Пусть для ϕ и ψ сложностью не более n справедливо равенство:

$$\frac{\phi}{\psi}\psi = \frac{\psi}{\phi}\phi \tag{4}$$

Шаг индукции. Шаг индукции будет доказан в три этапа.

Этап 1. Пусть ψ имеет сложность не более n, а $\phi = \phi_2\phi_1$, где ϕ_1 имеет сложность 1, а ϕ_2 – n.

$$\frac{\phi}{\psi}\psi = \frac{\phi_2\phi_1}{\psi}\psi \overset{(1)}{=} \frac{\phi_2}{\psi/\phi_1}\frac{\phi_1}{\psi}\psi \overset{(4)}{=} \frac{\phi_2}{\psi/\phi_1}\frac{\psi}{\phi_1}\phi_1 \overset{(4)}{=}$$

$$= \frac{\psi/\phi_1}{\phi_2}\phi_2\phi_1 \overset{(2)}{=} \frac{\psi}{\phi_2\phi_1}\phi_2\phi_1 = \frac{\psi/\phi}{\phi}$$

Этап 2. Пусть ϕ имеет сложность не более n, а $\psi = \psi_2\psi_1$, где ψ_1 имеет сложность 1, а ψ_2 – n.

Повторяя рассуждения предыдущего этапа в обратном порядке получаем, $(\psi/\phi)\phi = (\phi/\psi)\psi$.

На данный момент мы доказали, что для ϕ и ψ, один из которых сложности $n+1$, а второй не более n утверждение верно. Тем самым, можно считать, что мы расширили предположение индукции. Осталось завершить доказательство, показав, что в случае, когда оба преобразования имеют сложность $n+1$, утверждение также верно.

Этап 3. Пусть ψ имеет сложность не более $n+1$, а $\phi = \phi_2\phi_1$, где ϕ_1 имеет сложность 1, а ϕ_2 – n.

$$\frac{\phi}{\psi}\psi = \frac{\phi_2\phi_1}{\psi}\psi \overset{(1)}{=} \frac{\phi_2}{\psi/\phi_1}\frac{\phi_1}{\psi}\psi = \frac{\phi_2}{\psi/\phi_1}\frac{\psi}{\phi_1}\phi_1 = \frac{\psi/\phi_1}{\phi_2}\phi_2\phi_1 \overset{(2)}{=} \frac{\psi}{\phi_2\phi_1}\phi_2\phi_1 = \frac{\psi}{\phi}\phi$$

Следствие 6.1. *Синтаксический анализатор $ct\langle\frac{\phi}{\psi}\psi(G)\rangle$ совместим с синтаксическим анализатором $ct\langle\phi(G)\rangle$.*

Доказательство. Согласно Теореме 2, $ct\langle(\phi/\psi)\psi(G)\rangle = ct\langle(\psi/\phi)\phi(G)\rangle$ совместим с $ct\langle\phi(G)\rangle$, так как первое – это расширение второго.

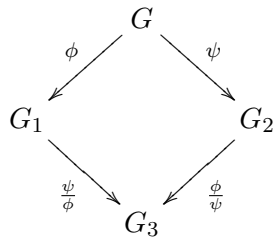
Сейчас можно дать ответы на все вопросы, сформулированные в начале работы.

Во-первых, операция $\frac{\psi}{\phi}$ – это и есть тот искомый способ модификации преобразования ψ так, чтобы его можно было применять после другого преобразования ϕ.

Во-вторых, главное требование к нему – это совместимость синтаксического анализатора для грамматики $\frac{\psi}{\phi}\phi(G)$ с синтаксическим анализатором грамматики $\phi(G)$. Это требование выполняется согласно следствию из теоремы.

В-третьих, как следует из теоремы, порядок применения преобразований значения не имеет.

Полученные результаты можно проиллюстрировать следующей диаграммой:



7. Заключение

Введённая операция ϕ/ψ меняет преобразование ϕ так, чтобы его можно было применять после преобразования ψ. Теорема же показывает, что в случае двух преобразований, не важно какое из них применять первым – результат будет одинаковым.

Эти выкладки фактически открывают дорогу к конструированию синтаксических анализаторов довольно сложных языков из независимых друг от друга простых частей. Каждая такая часть – это расширяющее преобразование исходной грамматики.