It's a Small Inverse

Kazuhiro Inaba (kinaba@nii.ac.jp)

September 10, 2010

1 Existence of Linear-Size Inverse

We assume the reader to be very familiar with the theory of tree transducers [Eng77, Bak79, EV85]. The goal of this note is the following theorem.

Theorem 1.1. *Let $f \in \mathrm{D_tMTT}^*$, i.e., f is a tree-to-tree function that can be expressed by a finite composition of total deterministic macro tree transducers. Then,*

$$\exists c \in \mathbb{N}.\ \forall t \in \mathrm{range}(f).\ \exists s \in \mathrm{dom}(f).\ \Big(f(s) = t \ \wedge\ |s| \leq c|t|\Big)$$

where $|s|$ and $|t|$ denotes the number of nodes in s and t, respectively.

We denote by ∃LBI (existentially linearly-bounded-input) the class of translations satisfying the property in the theorem. Hence, the theorem statement can be written as follows: $\mathrm{D_tMTT}^* \subseteq \exists\mathrm{LBI}$.

Let us take a look at an example. Consider the following translation f

$$\begin{aligned} f(\mathsf{a}(x_1)) &= f(x_1) \\ f(\mathsf{b}(x_1, x_2)) &= \mathsf{c}(f(x_1)) \\ f(\mathsf{e}) &= \mathsf{d} \end{aligned}$$

and the tree $t = \mathsf{c}(\mathsf{d})$. What is the input tree s that makes f to output t? There are many possibilities, because f is not injective. Some of them are large (e.g., $\mathsf{a(a(a(a(a(a(a(a(a(a(b(a(a(a(a(a(e))))),e)))))))))))}$), and some are not. In fact, the smallest one $\mathsf{b(e,e)}$ is *not so large* compared to the given tree t. More precisely speaking, the size is $\leq 2|t|$. The theorem tells us that this is a universal phenomenon shared among all MTT-definable translations.

1.1 The Proof

The problem is broken down to simpler cases by using the following lemma.

Lemma 1.2 ([Man02], Theorem 12)**.** $\mathrm{D_tMTT}^* \subseteq \mathrm{D_tLT^R}\,;\mathrm{LBI}$

X ; Y denotes sequential composition: X followed by Y. $\mathrm{D_tLT^R}$ is the class of total deterministic linear top-down tree transducers with regular lookahead, whose property is discussed soon. LBI is the class of translations that satisfies: $\exists c \in \mathbb{N}.\ \forall s \in \mathrm{dom}(f).\ |s| \leq c|f(s)|$. The class LBI is a strict subclass of ∃LBI.

Let us assume here that $\mathrm{D_tLT^R} \subseteq \exists\mathrm{LBI}$, then from Lemma 1.2 we can derive the main theorem, because we can also show the following fact.

Lemma 1.3. $\exists\mathrm{LBI}\,;\mathrm{LBI} \subseteq \exists\mathrm{LBI}$[1].

[1] For those who might wonder: it also holds that $\mathrm{LBI}\,;\mathrm{LBI} \subseteq \mathrm{LBI}$. But, $\exists\mathrm{LBI}\,;\exists\mathrm{LBI} \not\subseteq \exists\mathrm{LBI}$ and $\mathrm{LBI}\,;\exists\mathrm{LBI} \not\subseteq \exists\mathrm{LBI}$.

Proof. Let $g_1 \in \exists\text{LBI}$, $g_2 \in \text{LBI}$, and $t \in \text{range}(g_1\,;g_2)$. Then take arbitrary s_0, s_1 satisfying $s_1 = g_1(s_0)$ and $t = g_2(s_1)$ (since $t \in \text{range}(g_1\,;g_2)$, there exists at lease one such pair). From the assumption $g_1 \in \exists\text{LBI}$, we can make s_0 small. That is, by $s_1 \in \text{range}(g_1)$, there exists s'_0 s.t. $|s'_0| \leq c_{g_1}|s_1|$. By the LBI-property, $|s_1| \leq c_{g_2}|t|$. Hence, we can choose a small input s'_0 such that $|s'_0| \leq c_{g_1}c_{g_2}|t|$, as desired. □

Now, as you expect, the assumption on $\text{D}_\text{t}\text{LT}^\text{R}$ does hold.

Lemma 1.4. $\text{D}_\text{t}\text{LT}^\text{R} \subseteq \exists\text{LBI}$.

Proof. By Theorem 2.6 of [Eng77], $\text{D}_\text{t}\text{LT}^\text{R}$ can be represented as a deterministic finite-state bottom-up relabeling followed by $\text{D}_\text{t}\text{LT}$. Let g be the relabeling and $f \in \text{D}_\text{t}\text{LT}$, and t a tree in the range of $g\,;f$. Let s be (one of) the minimum input tree such that $(g\,;f)(s) = t$. We will show $|s|$ is bounded by $c|t|$ where c is a constant determined by g and f and independent from t or s.

For each node v of s, we denote by $f_\text{Q}(v)$ the state of f applied to v during the computation of $f(s)$ (since f is linear transducer, the state, if any, is uniquely determined; if f never visited v, let $f_q(v) = \bot$). Note that if $f_\text{Q}(v) = \bot$ then for all nodes v' in the subtree rooted at v, we have $f_\text{Q}(v') = \bot$. We can show $|\{v \mid f_\text{Q}(v) = \bot\}| \leq r|g|^r|\{v \mid f_\text{Q}(v) \neq \bot\}|$ where r is the maximum rank of the label alphabet, and $|g|$ is the number of states of the relabeling g. In other words, unvisited parts are smaller than visited parts (ignoring the constant factor).

The inequation is derived as follows. Let $v_1, \ldots, v_u$ be the set of nodes that they are unvisited ($f_\text{Q}(v_i) = \bot$) but all their ancestors are visited. It should be clear that $u \leq r|\{v \mid f_\text{Q}(v) \neq \bot\}|$; there can be at most $|\{v \mid f_\text{Q}(v) \neq \bot\}|$ leaves in the visited fragment of s, and each of them can only have at most r unvisited children. Furthermore, the number of nodes of each subtree rooted at v_i is bounded by $|g|^r$. Since the subtree of v_i is unvisited, we can freely substitute the subtree to another one without changing the output t, as long as the bottom-up relabeling g reaches the same state at v_i. Here, for any $|g|$-state tree automaton of rank r, its minimal instance tree is at most $|g|^r$ (by the pumping lemma of regular languages, the height of minimal instance is bounded by $|g|$). Hence we have the desired inequation.

We now know that the $\frac{1}{1+r|g|^r}$ fraction of the input tree s is visited by the translation f. Then we have $\frac{|s|}{1+r|g|^r} \leq |t|$, and the proof is done, isn't it? Unfortunately, no. Even if $\text{D}_\text{t}\text{LT}$ visits an input node, it does not mean that an output node is generated there. In the only one exceptional case, the transducer may *skip* the node without generating any new node. This happens when a rule of the form $q(\sigma(x_1, \ldots, x_m)) = q'(x_i)$ (the form whose rhs is a single state-application) is used at the node. We have to deal with this case.

Let V be the set of *skipped* nodes of s, i.e., the nodes v such that the right-hand side of the rule of f of for the state-label pair $f_\text{Q}(v), \text{label}(v)$ is a single state-application. Let $v_1, v_2, \ldots v_u \in V$ be a list of nodes such that v_i is the parent node of v_{i+1} for each i. Then, u must be less than or equal to $|f||g|$. This is again due to the pumping lemma. If the chain is longer than $|f||g|$, for some $i < j$ it would be $f_\text{Q}(v_i) = f_\text{Q}(v_j)$ and $g_\text{Q}(v_i) = g_\text{Q}(v_j)$ (where g_Q is the unique state used during the run of g on s), which can be shortened and hence contradicts the minimality of s.

This upper bound of the length implies that at least 1 out of $1 + |f||g|$ visited nodes are not skipped and generate some output node. In a summary, we have $|s| \leq (1 + r|g|^r)(1 + |f||g|)|t|$. □

1.2 Nondeterministic Version

I believe the same property holds for nondeterministic MTTs. Analogue of Lemma 1.2 also holds for the nondeterministic case (Theorem 5.10 of [Ina09]). Hence, all I have to do is to show $\text{LT}^\text{R} \subset \exists\text{LBI}$. Its proof should also be similar; after we fix one particular run of the transducer that converts s into t, the same argument should hold.

2 Discussion

Are there any easier proof for the theorem? The theorem may be trivial, just I'm not noticing it...

Lemma 1.4 may have something to do with Theorem 5.2 of [AU71], which (if I understand correctly) says that the growth rate of a single top-down tree transducer must be in the form x^c or c^x for some integer c. In other words, if the growth rate is less than linear, it is constant (no $\log n$ or $\sqrt{n}$ translation).

Finally, can we derive something fruitful from the theorem, like linear-time upper bound for some useful problem? Currently I have no idea... :p

References

[AU71] Alfred V. Aho and Jeffrey D. Ullman. Translations on a context-free grammar. *Information and Control*, 19:439–475, 1971.

[Bak79] Brenda S. Baker. Composition of top-down and bottom-up tree transductions. *Information and Control*, 41:186–213, 1979.

[Eng77] Joost Engelfriet. Top-down tree transducers with regular look-ahead. *Mathematical Systems Theory*, 10:289–303, 1977.

[EV85] Joost Engelfriet and Heiko Vogler. Macro tree transducers. *Journal of Computer and System Sciences*, 31:71–146, 1985.

[Ina09] Kazuhiro Inaba. *Complexity and Expressiveness of Models of XML Translations*. PhD thesis, The University of Tokyo, 2009.

[Man02] Sebastian Maneth. The complexity of compositions of deterministic tree transducers. In *Foundations of Software Technology and Theoretical Computer Science (FSTTCS)*, pages 265–276, 2002.