Learning XML : VPAs and Discrimination Trees

Cinzia Di Giusto, Davide Fissore, Etienne Lozes
Université Nice Côte d’Azur, CNRS, I3S, Sophia Antipolis, France
Stage d’été 2022

Why VPA?
For \( \forall \) Non-Deterministic VPA \( V_1 \), there \( \exists \) a Deterministic VPA \( V_2 \) such that \( L(V_1) = L(V_2) \) → Every binary operation between 2 VPA is decidable!

Note:
Push symbols ⇔ Open tags
Pop symbols ⇔ Close tags

Acceptance for XML:
Empty stack + final states

And Communication?
Arthur : Does \( w \in U \)?
Merlin : Yes/No

Arthur creates a conjecture \( C \).
Arthur : Does \( C = U \)?
Merlin : If \( C = U \) → Yes
else → a counter-example

What is Learning?
Dana Angluin’s framework :
The Learner wants to learn a language \( U \)
The Teacher knows \( U \)

<XML>
Text</XML>

XML:
XML (eXtensible Markup Language) is a standard format for data exchange.
XML representable w/VPA!

k-SEVPA
Single entry VPA are VPAs where states are partitioned into \( k \) modules.
Each module has only one entry for call transitions

«Canonical» VPA
Regular automata have a unique minimal (or canonical) representative, this is not true for VPA

ΔM(Σ)
It is a couple of words called well-matched words, \( u_i, u_j \), such that every call symbol of \( u = u_i \ldots u_j \) has a corresponding ret symbol

VPAs
VPA := Visibly pushdown automata. They can recognize context free languages.
The alphabet is :
\[ \Sigma = \Sigma_{\text{call}} \cup \Sigma_{\text{ret}} \cup \Sigma_{\text{int}} \]

An XML grammar to LEARN
\[
\begin{align*}
G := \\
d(\text{XML}) &= \text{Text} + \text{DIV} \\
d(\text{DIV}) &= \text{Text} + \text{DIV} \\
d : x \rightarrow \langle x \rangle \text{ RULE} \langle /x \rangle
\end{align*}
\]

Example:
\( \langle \text{XML} \rangle \langle \text{DIV} \rangle \langle \text{Text} \rangle \langle \text{DIV} \rangle \langle /\text{XML} \rangle \in G \)

Discrimination Tree
Thanks to Well-Matched words, we can build the Discrimination tree :
\( \bullet \) Inner Nodes contain a couple \( (u_1, u_2) \) forming a WM
\( \bullet \) Leaves are labelled with a string.

VPA from Disc. Tree?
From this discriminator tree, we can build the same VPA for the grammar \( G \).

<XML>Text</XML>

References
This work has been supported by the French government, through the UCA DS4H Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-17-EURE-0004.