A Coalgebraic View on Context-Free Languages

Joost Winter, Marcello Bonsangue, Jan Rutten

Centrum Wiskunde & Informatica

August 31, 2011

Overview

- The coalgebraic picture of regular languages and expressions is well-known.
- It is interesting to see how this work can be extended to context-free languages.
- We will give a coalgebraic treatment of context-free languages through systems of behavioural differential equations.
- Furthermore, another coalgebraic approach to context-free languages, through μ-expressions is given.

The $2 \times (-)^A$-functor

We will be concerned with coalgebras $f : X \to 2 \times X^A$.
We will use the following notation:

- Given $x \in X$, $o(x)$ (called the *output value of* x) will be the first component of $f(x)$.
- Given $x \in X$, x_a (called the *a-derivative of* x) will be the second component of $f(x)$, applied to a.

So: for every $x \in X$, $o(x)$ is an element of the set $\{0, 1\}$, and for every x and a, x_a is an element of X again.

Word derivatives

We can extend the notion of derivatives from alphabet symbols (i.e. elements of A) to words (i.e. elements of A^*) inductively:

- $x_\lambda = x$
- $x_{a \cdot w} = (x_a)_w$.

Homomorphisms and bisimulations (1)

Given two coalgebras (X, f) and (Y, g), a function $h : X \to Y$ is a *homomorphism* if the following hold:

1. For every $x \in X$, $o(x) = o(h(x))$.
2. For every $x \in X$ and $a \in A$, $h(x_a) = (h(x))_a$.

Homomorphisms and bisimulations (2)

Given two coalgebras (X, f) and (Y, g), a relation $R \subseteq X \times Y$ is a *bisimulation* if the following hold:

1. If $(x, y) \in R$, then $o(x) = o(y)$.
2. If $(x, y) \in R$, then for all $a \in A$, $(x_a, y_a) \in R$.

Final coalgebras

Consider the coalgebra $(\mathcal{L}, l)$ defined as follows:

- $\mathcal{L}$ is the set of all languages on the alphabet A.
- For any $L \in \mathcal{L}$:
  - $o(L)$ is 1 iff the empty word is in L.
  - $F_a = \{w \mid a \cdot w \in L\}$.

This is a *final coalgebra*: for every coalgebra (X, f), there is a *unique* homomorphism h from (X, f) to $(\mathcal{L}, l)$.
Given a coalgebra (X, f), and an element $x \in X$, we let $[\![x]\!]$ denote the value of x under this unique homomorphism.

The coalgebra of regular expressions

The set $\mathcal{E}$ of *regular expressions* over a finite alphabet A and the semiring $(2, +, \cdot, 0, 1)$ can be defined as follows:

$$t ::= a \in A \,|\, 0 \,|\, 1 \,|\, t + t \,|\, t \cdot t \,|\, t^*$$

We can assign a coalgebra structure to this set of regular expressions by specifying the output values and derivatives for each expression, giving us a coalgebra $(\mathcal{E}, e)$:

t	$o(t)$	t_a
$r \in \{0, 1\}$	r	0
$b \in A$	0	if $b = a$ then 1 else 0
$u + v$	$o(u) + o(v)$	$u_a + v_a$
$u \cdot v$	$o(u) \cdot o(v)$	$u_a \cdot v + o(u) \cdot v_a$
u^*	1	$u_a \cdot u^*$

Kleene's theorem, coalgebraically

- For any language $L \in \mathcal{L}$, we define the subcoalgebra generated by $\mathcal{L}$ as

$$\langle L \rangle := \{ L_w \mid w \in A^* \}$$

  This indeed generates a subcoalgebra, as $\langle L \rangle$ is closed under taking derivatives.

- Kleene's theorem, coalgebraically (Rutten, 1998): *For any $L \in \mathcal{L}$, $\langle L \rangle$ is finite iff there is a regular expression t such that $L = [\![t]\!]$.*

Introduction: context-free grammars and languages

- Going a step upward from the regular languages, in the Chomsky hierarchy, we arrive at the context-free languages.
- We present a format of coinductively defined systems of equations, which chracterize precisely the context-free languages.

Systems of equations (1)

We define the class TX of terms t as follows:

$$t ::= a \in A \,|\, x \in X \,|\, 0 \,|\, 1 \,|\, t + t \,|\, t \cdot t$$

where X is a finite set of variables, and A, as before, is a finite alphabet.
A well-formed system of equations consists of:

1. For every $x \in X$, exactly one equation of the form $o(x) = v$, where $v \in \{0, 1\}$.
2. For every $x \in X$ and $a \in A$, exactly one equation of the form $x_a = t$, where $t \in TX$.

Systems of equations (2)

Alternatively, we can consider a well-formed system of equations as a mapping

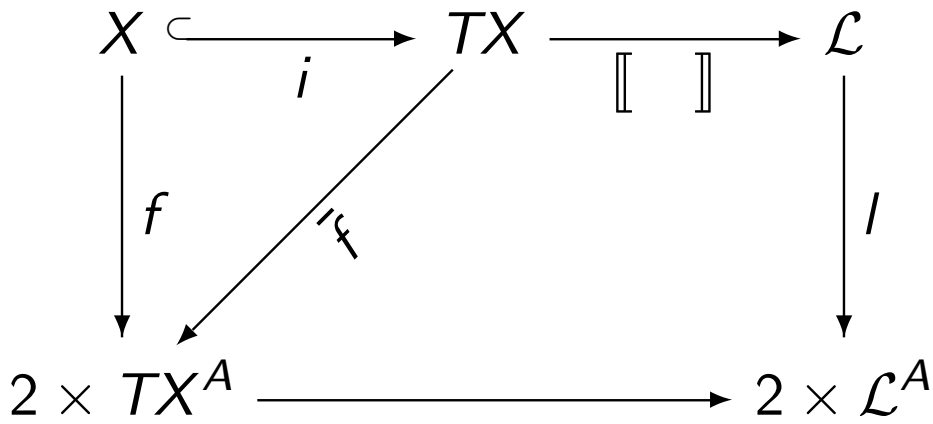
$$f : X \to 2 \times TX^A$$

We can extend such a mapping f to the $\mathcal{D}$-coalgebra $(TX, \bar{f})$ *generated by* (X, f) as follows:

t	$o(t)$	t_a
$x \in X$	$o(x)$	x_a (as specified by f)
$r \in \{0, 1\}$	r	0
$b \in A$	0	if $b = a$ then 1 else 0
$u + v$	$o(u) + o(v)$	$u_a + v_a$
$u \cdot v$	$o(u) \cdot o(v)$	$u_a \cdot v + o(u) \cdot v_a$

Systems of equations (3)

- This construction can be summarized diagrammatically:



- Proposition: *A language L is context-free iff there is a well-formed system of equations (X, f) and an $x \in X$, such that $[\![x]\!] = L$ w.r.t. the coalgebra $(TX, \bar{f})$ generated by it.*

From CFGs to systems of equations (1)

- We say a context-free grammar is in weak Greibach normal form, if every production rule has a right hand side either equal to the empty word λ, or of the form $a \cdot t$.
- Every CFG can be represented in weak Greibach normal form.

From CFGs to systems of equations (2)

We transform a CFG G in weak Greibach normal form into a system of equations as follows:

- We let the set X of variables be equal to the set of nonterminals in the grammar.
- Given a $x \in X$, we set $o(x) = 1$ iff the grammar contains a production rule $x \longrightarrow \lambda$.
- Given a $x \in X$ and an $a \in A$, we set

$$x_a = \sum \{w \mid x \longrightarrow a \cdot w\}$$

Given an initial symbol $x_0 \in X$, we now have $o((x_0)_w) = 1$ (and, hence, $w \in [\![x_0]\!]$) iff w is in the language generated by G.

From CFGs to systems of equations (3)

Example: Take the grammar:

$$x \to axa\,, \quad x \to ayb\,, \quad x \to aa\,, \quad y \to ayb\,, \quad y \to \lambda\,.$$

with $L(x) = \{a^{m+n}b^m a^n \mid m+n \geq 1\}$.
We obtain the following system of equations:

$$\begin{array}{lll} o(x) = 0 & x_a = \{xa, yb, a\} & x_b = \emptyset \\ o(y) = 1 & y_a = \{yb\} & y_b = \emptyset\,. \end{array}$$

From systems of equations to CFGs

Conversely, given a system of equations, we can construct a CFG in weak Greibach normal form:

- We first transform the system of equations to a new, equivalent, system, in which all derivatives are in disjunctive normal form, and do not contain any superfluous 0s or 1s.
- Derivatives in this new system are disjunctions of sequences of alphabet symbols and variables.
- We let the grammar include a rule $x \longrightarrow \lambda$ whenever $o(x) = 1$.
- We let the grammar include a rule $x \longrightarrow a \cdot w$, whenever w is a sequence of alphabet symbols and variables occurring as a disjunct in x_a.

μ-expressions (1)

Consider the class of μ-expressions t, and the class of *guarded* μ-expressions g, defined as follows:

$$\begin{aligned} t &::= 0 \mid 1 \mid x \in X \mid a \in A \mid t + t \mid t \cdot t \mid \mu x.g \\ g &::= 0 \mid 1 \mid a \cdot t\,(a \in A) \mid g + g \end{aligned}$$

μ-expressions (2)

We now define a coalgebra structure on the class of all *closed* μ-expressions:

t	$o(t)$	t_a
$r \in \{0,1\}$	r	0
$b \in A$	0	if $b = a$ then 1 else 0
$u + v$	$o(u) + o(v)$	$u_a + v_a$
$u \cdot v$	$o(u) \cdot o(v)$	$u_a \cdot v + o(u) \cdot v_a$
$\mu x.u$	$o(u[\mu x.u/x])$	$(u[\mu x.u/x])_a$

Due to the guardedness conditions, this is a well-defined coalgebra.

μ-expressions (3)

Proposition: *A language L is context-free iff there is a μ-expression t such that $[\![t]\!] = L$.*

Generalizing context-freeness

- Note that most of the definitions in this presentation do not necessarily require the underlying semiring to be the Boolean semiring!
- Taking the definition of *regular expressions* and applying it to arbitrary semirings, we obtain rational streams and power series, which are well-known.
- Furthermore, we can easily generalize the notion above, of *context-free languages*, to *context-free streams* and *context-free power series*.

Context-free streams

Example: When we take the semiring of natural numbers as underlying semiring, the Catalan numbers

$$1, 1, 2, 5, 14, 42, 132, 429, 1430, \ldots$$

are context-free, and are generated by the following system of equations:

$$o(x) = 1 \quad x' = x \cdot x$$

Conclusions and further work

- There is a very neat coalgebraic representation of regular expressions, and Kleene's theorem can be expressed succinctly in a coalgebraic fashion.
- We have extended this work towards context-free languages and grammars, and provided a coalgebraic characterization using systems of equations.
- The first steps towards a generalization to other functors of the type $R \times (-)^A$ havs been made, and there are some nice examples of context-free streams.
- Future work: further investigate this notion of 'generalized context-freeness' of power series, and see how this relates to other, existing notions.