

עיבוד שפות טבעיות

שולי וינטנר

Introduction to XFST

Introduction to XFST

- XFST is an interface giving access to finite-state operations (algorithms such as union, concatenation, iteration, intersection, composition etc.)
- XFST includes a regular expression compiler
- The interface of XFST includes a lookup operation (*apply up*) and a generation operation (*apply down*)
- The regular expression language employed by XFST is an extended version of standard regular expressions

Introduction to XFST

Introduction to XFST

a a simple symbol

Introduction to XFST

a a simple symbol

c a t a concatenation of three symbols

Introduction to XFST

- a a simple symbol
- c a t a concatenation of three symbols
- [c a t] grouping brackets

Introduction to XFST

- a a simple symbol
- c a t a concatenation of three symbols
- [c a t] grouping brackets
- ? denotes any single symbol

Introduction to XFST

- a a simple symbol
- c a t a concatenation of three symbols
- [c a t] grouping brackets
- ? denotes any single symbol
- ' '+Noun' ' single symbol with multicharacter print name

Introduction to XFST

a	a simple symbol
c a t	a concatenation of three symbols
[c a t]	grouping brackets
?	denotes any single symbol
' '+Noun'	single symbol with multicharacter print name
%+Noun	single symbol with multicharacter print name

Introduction to XFST

a	a simple symbol
c a t	a concatenation of three symbols
[c a t]	grouping brackets
?	denotes any single symbol
' '+Noun'	single symbol with multicharacter print name
%+Noun	single symbol with multicharacter print name
cat	a single multicharacter symbol

Introduction to XFST

a	a simple symbol
c a t	a concatenation of three symbols
[c a t]	grouping brackets
?	denotes any single symbol
' '+Noun'	single symbol with multicharacter print name
%+Noun	single symbol with multicharacter print name
cat	a single multicharacter symbol
{cat}	equivalent to [c a t]

Introduction to XFST

[] the empty string

Introduction to XFST

[] the empty string

0 the empty string

Introduction to XFST

- [] the empty string
- 0 the empty string
- [A] bracketing; equivalent to A

Introduction to XFST

[]	the empty string
0	the empty string
[A]	bracketing; equivalent to A
A B	union

Introduction to XFST

[]	the empty string
0	the empty string
[A]	bracketing; equivalent to A
A B	union
(A)	optionality; equivalent to [A 0]

Introduction to XFST

[]	the empty string
0	the empty string
[A]	bracketing; equivalent to A
A B	union
(A)	optionality; equivalent to [A 0]
A&B	intersection

Introduction to XFST

[]	the empty string
0	the empty string
[A]	bracketing; equivalent to A
A B	union
(A)	optionality; equivalent to [A 0]
A&B	intersection
A B	concatenation

Introduction to XFST

[]	the empty string
0	the empty string
[A]	bracketing; equivalent to A
A B	union
(A)	optionality; equivalent to [A 0]
A&B	intersection
A B	concatenation
A-B	set difference

Introduction to XFST

A*

Kleene-star

Introduction to XFST

A* Kleene-star

A+ one or more iterations

Introduction to XFST

- A* Kleene-star
- A+ one or more iterations
- ?* the universal language

Introduction to XFST

A^*	Kleene-star
A^+	one or more iterations
$?^*$	the universal language
$\sim A$	the complement of A ; equivalent to $[?^* - A]$

Introduction to XFST

A^*	Kleene-star
A^+	one or more iterations
$?^*$	the universal language
$\sim A$	the complement of A ; equivalent to $[?^* - A]$
$\sim [?^*]$	the empty language

Introduction to XFST

A^*	Kleene-star
A^+	one or more iterations
$?^*$	the universal language
$\sim A$	the complement of A ; equivalent to $[?^* - A]$
$\sim [?^*]$	the empty language
$\%+$	the literal plus-sign symbol

Introduction to XFST

A^*	Kleene-star
A^+	one or more iterations
$?^*$	the universal language
$\sim A$	the complement of A ; equivalent to $[?^* - A]$
$\sim [?^*]$	the empty language
$\%+$	the literal plus-sign symbol
$\%^*$	the literal asterisk symbol (and similarly for $\%?$, $\%($, $\%]$ etc.

Introduction to XFST – denoting relations

Introduction to XFST – denoting relations

A .x. B Cartesian product; relates every string in A to every string in B

Introduction to XFST – denoting relations

$A \ .x.\ B$ Cartesian product; relates every string in A to every string in B

$a:b$ shorthand for $[a \ .x.\ b]$

Introduction to XFST – denoting relations

- A .x. B Cartesian product; relates every string in A to every string in B
- a:b shorthand for [a .x. b]
- %+P1:s shorthand for [%+P1 .x. s]

Introduction to XFST – denoting relations

$A \ .x.\ B$ Cartesian product; relates every string in A to every string in B

$a:b$ shorthand for $[a \ .x.\ b]$

$\%+Pl:s$ shorthand for $[\%+Pl \ .x.\ s]$

$\%+Past:ed$ shorthand for $[\%+Past \ .x.\ ed]$

Introduction to XFST – denoting relations

$A \ .x.\ B$ Cartesian product; relates every string in A to every string in B

$a:b$ shorthand for $[a \ .x.\ b]$

$\%+Pl:s$ shorthand for $[\%+Pl \ .x.\ s]$

$\%+Past:ed$ shorthand for $[\%+Past \ .x.\ ed]$

$\%+Prog:ing$ shorthand for $[\%+Prog \ .x.\ ing]$

Introduction to XFST – useful abbreviations

Introduction to XFST – useful abbreviations

$\$A$ the language of all the strings that contain A;
equivalent to $[?* A ?*]$

Introduction to XFST – useful abbreviations

- $\$A$ the language of all the strings that contain A ;
equivalent to $[?* A ?*]$
- A/B the language of all the strings in A , ignoring any
strings from B , e.g.,

Introduction to XFST – useful abbreviations

- $\$A$ the language of all the strings that contain A;
equivalent to $[?* A ?*]$
- A/B the language of all the strings in A, ignoring any
strings from B, e.g.,
- $a*/b$ includes strings such as a, aa, aaa, ba, ab,
aba etc.

Introduction to XFST – useful abbreviations

- $\$A$ the language of all the strings that contain A; equivalent to $[?* A ?*]$
- A/B the language of all the strings in A, ignoring any strings from B, e.g.,
- $a*/b$ includes strings such as a, aa, aaa, ba, ab, aba etc.
- $\setminus A$ any single symbol, minus strings in A. Equivalent to $[? - A]$, e.g.,

Introduction to XFST – useful abbreviations

- $\$A$ the language of all the strings that contain A ; equivalent to $[?* A ?*]$
- A/B the language of all the strings in A , ignoring any strings from B , e.g.,
- $a*/b$ includes strings such as a , aa , aaa , ba , ab , aba etc.
- $\setminus A$ any single symbol, minus strings in A . Equivalent to $[? - A]$, e.g.,
- $\setminus b$ any single symbol, except 'b'. Compare to:
- $\sim A$ the complement of A , i.e., $[?* - A]$

Introduction to XFST – example

Introduction to XFST – example

```
[ [l e a v e %+VBZ .x. l e a v e s] |  
[l e a v e %+VB .x. l e a v e] |  
[l e a v e %+VBG .x. l e a v i n g] |  
[l e a v e %+VBD .x. l e f t] |  
[l e a v e %+NN .x. l e a v e] |  
[l e a v e %+NNS .x. l e a v e s] |  
[l e a f %+NNS .x. l e a v e s ] |  
[l e f t %+JJ .x. l e f t] ]
```

Introduction to XFST – user interface

Introduction to XFST – user interface

```
prompt% H:\class\data\shuly\xfst
```

```
xfst> help
```

```
xfst> help union net
```

```
xfst> exit
```

```
xfst> read regex [d o g | c a t];
```

```
xfst> read regex < myfile.regex
```

```
xfst> apply up dog
```

```
xfst> apply down dog
```

```
xfst> pop stack
```

```
xfst> clear stack
```

```
xfst> save stack myfile.fsm
```

Introduction to XFST – example of lookup and generation

APPLY DOWN> leave+VBD

left

APPLY UP> leaves

leave+NNS

leave+VBZ

leaf+NNS

Introduction to XFST – variables

Introduction to XFST – variables

```
xfst> define Myvar;
```

```
xfst> define Myvar2 [d o g | c a t];
```

```
xfst> undefine Myvar;
```

```
xfst> define var1 [b i r d | f r o g | d o g];
```

```
xfst> define var2 [d o g | c a t];
```

```
xfst> define var3 var1 | var2;
```

```
xfst> define var4 var1 var2;
```

```
xfst> define var5 var1 & var2;
```

```
xfst> define var6 var1 - var2;
```

Introduction to XFST – variables

```
xfst> define Root [w a l k | t a l k | w o r k];  
xfst> define Prefix [0 | r e];  
xfst> define Suffix [0 | s | e d | i n g];  
xfst> read regex Prefix Root Suffix;  
xfst> words  
xfst> apply up walking
```

Introduction to XFST – replace rules

Replace rules are an extremely powerful extension of the regular expression metalanguage.

Introduction to XFST – replace rules

Replace rules are an extremely powerful extension of the regular expression metalanguage.

The simplest replace rule is of the form

$$\textit{upper} \rightarrow \textit{lower} \parallel \textit{leftcontext} _ \textit{rightcontext}$$

Its denotation is the relation which maps string to themselves, with the exception that an occurrence of *upper* in the input string, preceded by *leftcontext* and followed by *rightcontext*, is replaced in the output by *lower*.

Introduction to XFST – replace rules

The language Bambona has an underspecified nasal morpheme N that is realized as a labial m or as a dental n depending on its environment: N is realized as m before p and as n elsewhere.

The language also has an assimilation rule which changes p to m when the p is followed by m .

Introduction to XFST – replace rules

The language Bambona has an underspecified nasal morpheme *N* that is realized as a labial *m* or as a dental *n* depending on its environment: *N* is realized as *m* before *p* and as *n* elsewhere.

The language also has an assimilation rule which changes *p* to *m* when the *p* is followed by *m*.

```
xfst> clear stack ;  
xfst> define Rule1 N -> m || _ p ;  
xfst> define Rule2 N -> n ;  
xfst> define Rule3 p -> m || m _ ;  
xfst> read regex Rule1 .o. Rule2 .o. Rule3 ;
```

Introduction to XFST – replace rules

Word boundaries can be explicitly referred to:

```
xfst> define Vowel [a|e|i|o|u];  
xfst> e -> ' || [.#.] [c | d | l | s] _ [% Vowel];
```

Introduction to XFST – replace rules

Contexts can be omitted:

```
xfst> define Rule1 N -> m || _ p ;  
xfst> define Rule2 N -> n ;  
xfst> define Rule3 p -> m || m _ ;
```

This can be used to clear unnecessary symbols introduced for “bookkeeping”:

```
xfst> define Rule1 %^MorphmeBoundary -> 0;
```

Introduction to XFST – replace rules

Rules can define multiple replacements:

[A -> B, B -> A]

or multiple replacements that share the same context:

[A -> B, B -> A || L _ R]

or multiple contexts:

[A -> B || L1 _ R1, L2 _ R2]

or multiple replacements and multiple contexts:

[A -> B, B -> A || L1 _ R1, L2 _ R2]

Introduction to XFST – replace rules

Rules can apply in parallel:

```
xfst> clear stack
```

```
xfst> read regex a -> b .o. b -> a ;
```

```
xfst> apply down abba
```

```
aaaa
```

```
xfst> clear stack
```

```
xfst> read regex b -> a .o. a -> b ;
```

```
xfst> apply down abba
```

```
bbbb
```

```
xfst> clear stack
```

```
xfst> read regex a -> b , b -> a ;
```

```
xfst> apply down abba
```

```
baab
```

Introduction to XFST – replace rules

When rules that have contexts apply in parallel, the rule separator is a double comma:

```
xfst> clear stack
```

```
xfst> read regex
```

```
b -> a || .#. s ?* _ , , a -> b || _ ?* e .#. ;
```

```
xfst> apply down sabbae
```

```
sbaabe
```


Introduction to XFST – marking

The special symbol “...” in the right-hand side of a replace rule stands for whatever was matched in the left-hand side of the rule.

Introduction to XFST – marking

The special symbol “...” in the right-hand side of a replace rule stands for whatever was matched in the left-hand side of the rule.

```
xfst> clear stack;  
xfst> read regex [a|e|i|o|u] -> %[ ... ];  
xfst> apply down unnecessarily  
[u]nn[e]c[e]ss[a]r[i]ly
```

Introduction to XFST – marking

```
xfst> clear stack;
xfst> read regex [a|e|i|o|u]+ -> %[ ... %];
xfst> apply down feeling
f[e][e]l[i]ng
f[ee]l[i]ng
xfst> apply down poolcleaning
p[o][o]lcl[e][a]n[i]ng
p[oo]lcl[e][a]n[i]ng
p[o][o]lcl[ea]n[i]ng
p[oo]lcl[ea]n[i]ng
xfst> read regex [a|e|i|o|u]+ @-> %[ ... %];
xfst> apply down poolcleaning
p[oo]lcl[ea]n[i]ng
```

Introduction to XFST – shallow parsing

Introduction to XFST – shallow parsing

Assume that text is represented as strings of part-of-speech tags, using 'd' for determiner, 'a' for adjective, 'n' for noun, and 'v' verb, etc. In other words, in this example the regular expression symbols represent whole words rather than single letters in a text.

Introduction to XFST – shallow parsing

Assume that text is represented as strings of part-of-speech tags, using ‘d’ for determiner, ‘a’ for adjective, ‘n’ for noun, and ‘v’ verb, etc. In other words, in this example the regular expression symbols represent whole words rather than single letters in a text.

Assume that a noun phrase consists of an optional determiner, any number of adjectives, and one or more nouns:

$[(d) a^* n^+]$

This expression denotes an infinite set of strings, such as “n” (cats), “aan” (discriminating aristocratic cats), “nn” (cat food), “dn” (many cats), “dann” (that expensive cat food) etc.

Introduction to XFST – shallow parsing

A simple noun phrase parser can be thought of as a transducer that inserts markers, say, a pair of braces { }, around noun phrases in a text. The task is not as trivial as it seems at first glance. Consider the expression

```
[(d) a* n+ -> %{ ... %}]
```

Applied to the input “danvn” (many small cats like milk) this transducer yields three alternative bracketings:

```
xfst> apply down danvn
```

```
da{n}v{n}
```

```
d{an}v{n}
```

```
{dan}v{n}
```

Introduction to XFST – longest match

For certain applications it may be desirable to produce a unique parse, marking the maximal expansion of each NP: “{dan}v{n}”. Using the left-to-right, longest-match replace operator @-> instead of the simple replace operator -> yields the desired result:

```
[(d) a* n+ @-> % { ... %}]
```

```
xfst> apply down danvn  
{dan}v{n}
```

Introduction to XFST – the coke machine

Introduction to XFST – the coke machine

A vending machine dispenses drinks for 65 cents a can. It accepts any sequence of the following coins: 5 cents (represented as 'n'), 10 cents ('d') or 25 cents ('q'). Construct a regular expression that compiles into a finite-state automaton that implements the behavior of the soft drink machine, pairing "PLONK" with a legal sequence that amounts to 65 cents.

Introduction to XFST – the coke machine

The construction A^n denotes the concatenation of A with itself n times.

Introduction to XFST – the coke machine

The construction A^n denotes the concatenation of A with itself n times.

Thus the expression $[n .x. c^5]$ expresses the fact that a nickel is worth 5 cents.

Introduction to XFST – the coke machine

The construction A^n denotes the concatenation of A with itself n times.

Thus the expression $[n .x. c^5]$ expresses the fact that a nickel is worth 5 cents.

A mapping from all possible sequences of the three symbols to the corresponding value:

$$[[n .x. c^5] \mid [d .x. c^{10}] \mid [q .x. c^{25}]]^*$$

The solution:

$$[[n .x. c^5] \mid [d .x. c^{10}] \mid [q .x. c^{25}]]^*$$

.o.

$$[c^{65} .x. PLONK]$$

Introduction to XFST – the coke machine

```
clear stack
define SixtyFiveCents
[[n .x. c^5] | [d .x. c^10] | [q .x. c^25]]* ;
define BuyCoke
SixtyFiveCents .o. [c^65 .x. PLONK] ;
```

Introduction to XFST – the coke machine

In order to ensure that extra money is paid back, we need to modify the lower language of BuyCoke to make it a subset of $[PLONK^* q^* d^* n^*]$.

Introduction to XFST – the coke machine

In order to ensure that extra money is paid back, we need to modify the lower language of BuyCoke to make it a subset of $[PLONK^* q^* d^* n^*]$.

To ensure that the extra change is paid out only once, we need to make sure that quarters get paid before dimes and dimes before nickels.

Introduction to XFST – the coke machine

In order to ensure that extra money is paid back, we need to modify the lower language of BuyCoke to make it a subset of $[PLONK^* q^* d^* n^*]$.

To ensure that the extra change is paid out only once, we need to make sure that quarters get paid before dimes and dimes before nickels.

```
clear stack
define SixtyFiveCents
[[n .x. c5] | [d .x. c10] | [q .x. c25]]* ;
define ReturnChange SixtyFiveCents .o.
[[c65 .x. PLONK]* [c25 .x. q]*
[c10 .x. d]* [c5 .x. n]*] ;
```

Introduction to XFST – the coke machine

The next refinement is to ensure that as much money as possible is converted into soft drinks and to remove any ambiguity in how the extra change is to be reimbursed.

Introduction to XFST – the coke machine

The next refinement is to ensure that as much money as possible is converted into soft drinks and to remove any ambiguity in how the extra change is to be reimbursed.

```
clear stack
define SixtyFiveCents
[[n .x. c^5] | [d .x. c^10] | [q .x. c^25]]* ;
define ReturnChange SixtyFiveCents .o.
[[c^65 .x. PLONK]* [c^25 .x. q]*
[c^10 .x. d]* [c^5 .x. n]*] ;
define ExactChange ReturnChange .o.
[~$[q q q | [q q | d] d [d | n] | n n]] ;
```

Introduction to XFST – the coke machine

To make the machine completely foolproof, we need one final improvement. Some clients may insert unwanted items into the machine (subway tokens, foreign coins, etc.). These objects should not be accepted; they should be passed right back to the client. This goal can be achieved easily by wrapping the entire expression inside an ignore operator.

Introduction to XFST – the coke machine

To make the machine completely foolproof, we need one final improvement. Some clients may insert unwanted items into the machine (subway tokens, foreign coins, etc.). These objects should not be accepted; they should be passed right back to the client. This goal can be achieved easily by wrapping the entire expression inside an ignore operator.

```
define IgnoreGarbage  
[ [ ExactChange ]/[\[q | d | n]]] ;
```