Approximate Equivalence of Transduction

CMF 2007

ADRIEN VIEILLERIBIÈRE

Laboratoire de Recherche en Informatique
Université Paris Sud XI

May 21st, 2007

Motivations

1. Data Exchange
2. xslt Transformations
3. Model versus Specification

Outline: Preliminaries

1. Word Edit Distance with Moves
2. Statistics
   - Uniform Statistic of a Word
   - Link Statistics-Distance
   - Uniform Statistic of a Loop
3. Word Transducer
   - Definition
   - Relation defined by a Transducer
   - Powering
   - Graph of the strongly connected components

Elementary operations

- Insert a letter : $aaabbb \xrightarrow{ins} aaa\mathbf{c}bbb$
- Remove a letter : $aaacbb\mathbf{b} \xrightarrow{del} aaacbb$
- Change a label : $aaacb\mathbf{b} \xrightarrow{swi} aaacbc$
- Move a subword : $a\mathbf{aa}cbc \xrightarrow{mov} acb\mathbf{aa}c$

Edit Distance with moves

1. $\mathsf{dist}(w_1, w_2) = \min\{p \mid w_1 \xrightarrow{o_1} \ldots \xrightarrow{o_p} w_2\}$
2. $\mathsf{d}(w_1, w_2) = \dfrac{\mathsf{dist}(w_1, w_2)}{\max(|w_1|, |w_2|)}$
3. $\mathsf{dist}(w_1, L) = \min_{w_2 \in L} \mathsf{dist}(w_1, w_2)$
4. $\mathsf{dist}(L_1, L_2) = \min_{w_1 \in L_2, w_2 \in L_2} \mathsf{dist}(w_1, w_2)$

Edit Distance with moves

1. $\mathsf{dist}(w_1, w_2) = \min\{p \mid w_1 \xrightarrow{o_1} \ldots \xrightarrow{o_p} w_2\}$
2. $\mathsf{d}(w_1, w_2) = \dfrac{\mathsf{dist}(w_1, w_2)}{\max(|w_1|, |w_2|)}$
3. $\mathsf{dist}(w_1, L) = \min_{w_2 \in L} \mathsf{dist}(w_1, w_2)$
4. $\mathsf{dist}(L_1, L_2) = \min_{w_1 \in L_2, w_2 \in L_2} \mathsf{dist}(w_1, w_2)$

Examples

1. $\mathsf{dist}(aaabbb, bbbaac) = 2$
2. $\mathsf{d}(aaabbb, bbbaac) = 1/3$
3. $\mathsf{dist}(aaabbb, a^*) = 3$
4. $\mathsf{dist}\left((aaa)^+.(bbbbbbbb)^+ \ , \ (bbbbbbbb)^*.(aaa)^+\right) = 1$

1. Word Edit Distance with Moves

2. Statistics
   - Uniform Statistic of a Word
   - Link Statistics-Distance
   - Uniform Statistic of a Loop

3. Word Transducer
   - Definition
   - Relation defined by a Transducer
   - Powering
   - Graph of the strongly connected components

Let Σ be a binary alphabet, $w \in \Sigma^*$ of length n and $k \in \mathbb{N}^+$

Statistic of a Word : ustat_k

$$\mathsf{ustat}_k(w) = \frac{1}{n-k+1}. \begin{pmatrix} \#n_1 \\ \#n_2 \\ \vdots \\ \#n_{2^k} \end{pmatrix} \text{ where } \begin{array}{c} \#n_1 : \text{number of } 00\ldots0 \\ \#n_2 : \text{number of } 0\ldots01 \\ \vdots \\ \#n_{2^k} : \text{number of } 1\ldots11 \end{array}$$

Example

$w = 001010101110$, length $n = 12$, $k = 2$, $n - k + 1 = 11$ blocks

$$\mathsf{ustat}(w) = \frac{1}{11} \begin{pmatrix} 1 \\ 4 \\ 4 \\ 2 \end{pmatrix}$$

Let Σ be a alphabet, $w \in \Sigma^*$ of length n and $k \in \mathbb{N}^+$

Statistic of a Word : ustat_k

$$\mathsf{ustat}_k(w) = \frac{1}{n-k+1} \cdot \begin{pmatrix} \#n_1 \\ \#n_2 \\ \vdots \\ \#n_{|\Sigma|^k} \end{pmatrix}$$

Main Properties

1. $\mathrm{d}(w, w') \approx |\mathsf{ustat}(w) - \mathsf{ustat}(w')|_1$ for words of similar length
2. $|\mathsf{ustat}(w) - Y(w)|_1 \leq \varepsilon$ for $Y(W)$ the ustat vector on N samples
3. distance between words
   - NP-complete
   - Testable, $\mathcal{O}(1)$:
     Sample N subwords of length k: $Y(w)$ and $Y(w')$.
     If $|Y(w) - Y(w')|_1 < \varepsilon = \frac{1}{k}$ accept, else reject

Limit Statistic of a loop

$$\text{Let } L = w^* \ , \ |w| = q$$

$$\mathcal{H}_{\text{lim}}(L)[u] = \Pr_{j=1,\dots,q}\left[w[j].w[j{+}1 \text{ Mod } k]\dots w[j{+}k{-}1 \text{ Mod } k] = u\right]$$

Example

$$L = (0001)^*, k = 2, \mathcal{H}_{\text{lim}}(L) = \begin{pmatrix} 1/2 \\ 1/4 \\ 1/4 \\ 0 \end{pmatrix}$$

1. Word Edit Distance with Moves

2. Statistics
   - Uniform Statistic of a Word
   - Link Statistics-Distance
   - Uniform Statistic of a Loop

3. Word Transducer
   - Definition
   - Relation defined by a Transducer
   - Powering
   - Graph of the strongly connected components

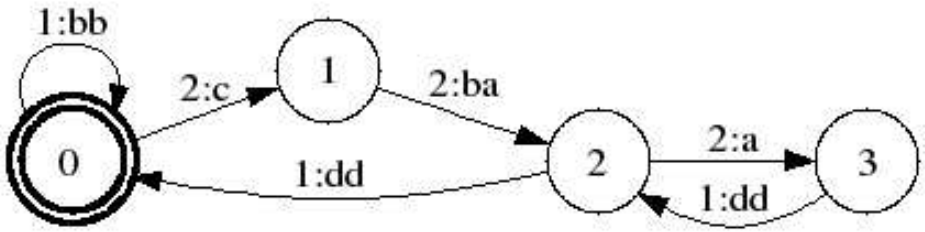
Word Transducers

A word transducter $\mathscr{T}$ is a tuple $(\Sigma, \Sigma', Q, I, F, \delta)$ where

- Σ in the input alphabet, Σ' the output alphabet
- Q a set of states, $I \in Q$ an initial state, $F \subseteq Q$ final states
- $\delta : Q \times (\Sigma \cup \Lambda) \times (\Sigma')^* \times Q$ the transition fonction

Example of a Word Transducer

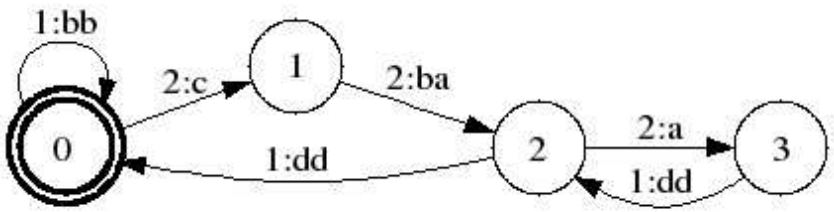
- $\Sigma = \{1, 2\}, \Sigma' = \{a, b, c, d\}$
- $Q = \{0, 1, 2, 3\}, I = \{0\}, F = \{0\}$
- $\delta = \{(0, 1, bb, 0), (0, 2, c, 1), (1, 2, ba, 2), (2, 1, dd, 0), (2, 2, a, 3), (3, 1, dd, 2)\}$



Relation of a Word Transducer $\mathscr{T}$

For two words w_1 and w_2, (w_1, w_2) is in the relation defined by $\mathscr{T}$, if there is a run from the initial state to a final state inputing w_1 and outputing w_2. [1]

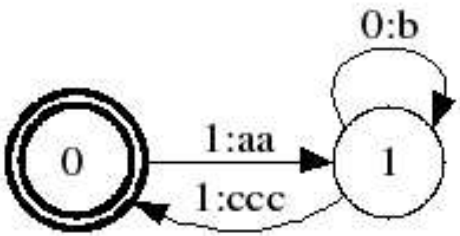
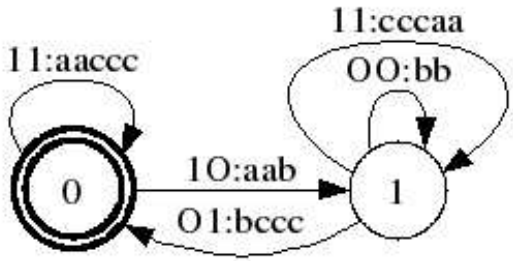
Example



$(\Lambda, \Lambda), (1, bb), (22211, cbaadddd)$ are in the relation defined by $\mathscr{T}$

[1] We then note $(w_1, w_2) \in \mathscr{T}$

Example : Squaring

a transducer	its square
	

k-Powering of a Transducer

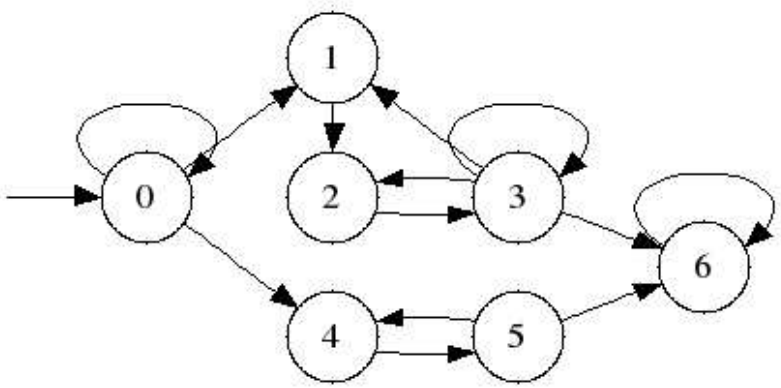
Let $\mathscr{T} = (\Sigma, \Sigma', Q, I, F, \delta)$ a transducer and $k \in \mathbb{N}$.

$$\mathscr{T}^k = (\Sigma^k, \Sigma', Q, I, F, \delta^k) \text{ where}$$

$(q_1\ ,\ a_1a_2...a_k\ ,\ w\ ,\ q_{k+1}) \in \delta^k$ iff there exist $q_1, ..., q_k \in Q$ and $w_1, ..., w_k \in (\Sigma')^*$ such that $w = w_1.w_2...w_k$ and for all i in $\{1, ..., k\}$, $(q_i, a_i, w_i, q_{i+1}) \in \delta$.

For G the graph of a (delabeled) automaton or transducer, $\widehat{G}$ denotes the graph of strongly connected components of G

Example

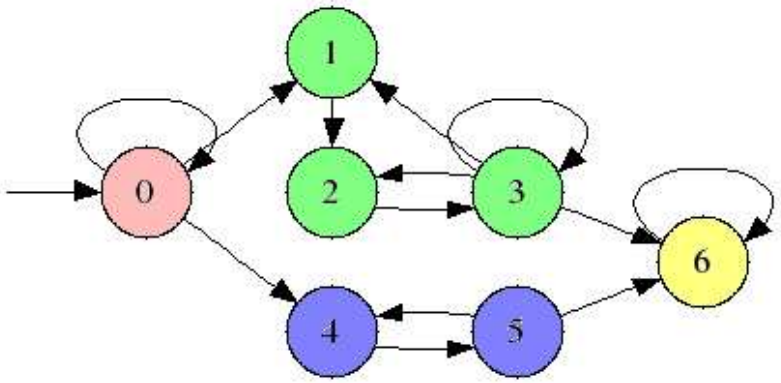


G

For G the graph of a (delabeled) automaton or transducer, $\widehat{G}$ denotes the graph of strongly connected components of G

Example

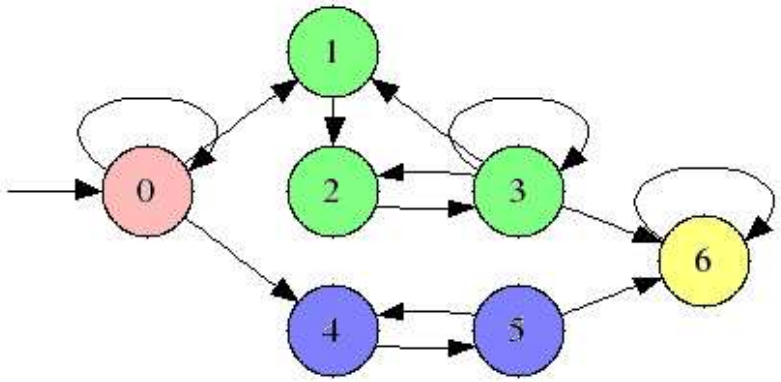
G



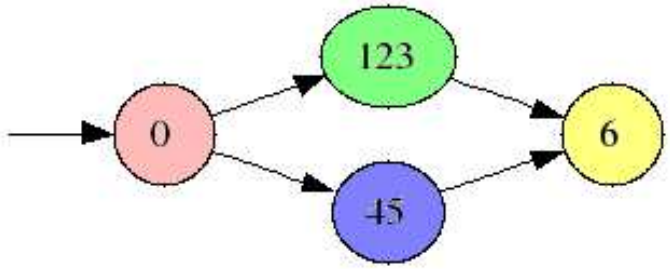
For G the graph of a (delabeled) automaton or transducer, $\widehat{G}$ denotes the graph of strongly connected components of G

Example

G



$\widehat{G}$



Outline: Equivalences

4 warm up

5 definition

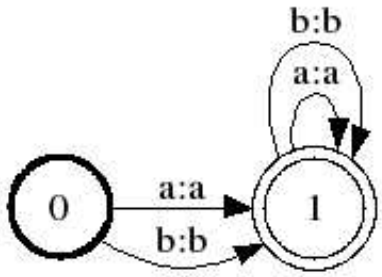
6 exact case

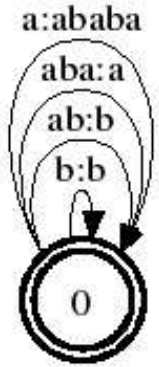
7 ε-equivalence
- two transducers ε-close
- statistical representation
- In the relation => close to the embedding
- In the embedding => close to the relation
- counterexamples : sources of non ε-equivalence

The intersection is undecidable

Reduction to the Post Correspondance Problem

- an instance: $\left(\begin{array}{c} a \\ ababa \end{array}\right), \left(\begin{array}{c} aba \\ a \end{array}\right), \left(\begin{array}{c} ab \\ b \end{array}\right), \left(\begin{array}{c} b \\ b \end{array}\right)$
- a solution: $\left(\begin{array}{c} aba \\ a \end{array}\right)\left(\begin{array}{c} b \\ b \end{array}\right)\left(\begin{array}{c} a \\ ababa \end{array}\right)\left(\begin{array}{c} b \\ b \end{array}\right)\left(\begin{array}{c} aba \\ a \end{array}\right)$

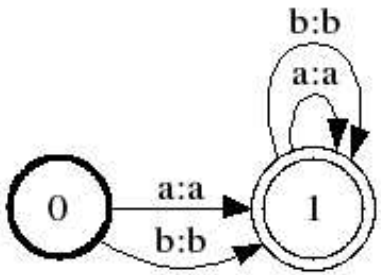


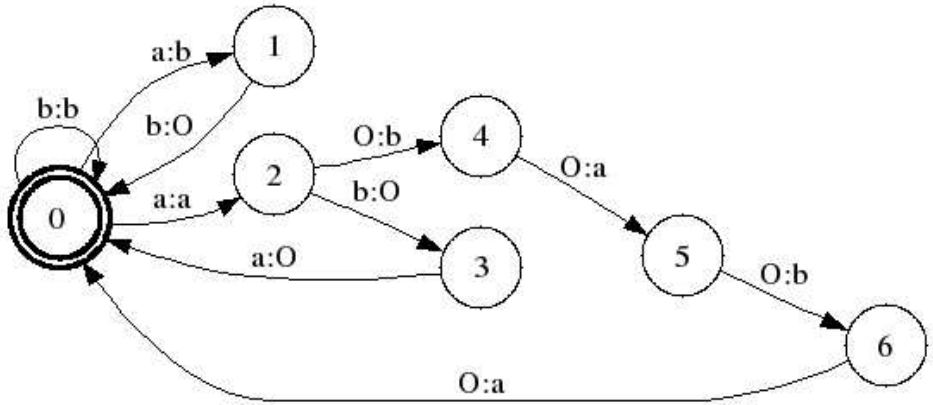


The intersection is undecidable

Reduction to the Post Correspondance Problem

- an instance: $\left(\begin{array}{c} a \\ ababa \end{array}\right), \left(\begin{array}{c} aba \\ a \end{array}\right), \left(\begin{array}{c} ab \\ b \end{array}\right), \left(\begin{array}{c} b \\ b \end{array}\right)$
- a solution: $\left(\begin{array}{c} aba \\ a \end{array}\right)\left(\begin{array}{c} b \\ b \end{array}\right)\left(\begin{array}{c} a \\ ababa \end{array}\right)\left(\begin{array}{c} b \\ b \end{array}\right)\left(\begin{array}{c} aba \\ a \end{array}\right)$





From $\mathscr{T}_1 \subseteq \mathscr{T}_2$ to $\mathscr{T}_1 \subseteq_\varepsilon \mathscr{T}_2$

1. **exact inclusion:** $(w, w') \in \mathscr{T}_1 \implies (w, w') \in \mathscr{T}_2$
2. **ε-inclusion:** $(w, w') \in \mathscr{T}_1 \implies \exists\, (\widetilde{w}, \widetilde{w'}) \in \mathscr{T}_2$ such that, for the classical edit distance with moves
   - $\widetilde{w}$ is ε-close to w *i.e.* $\mathrm{d}(w, \widetilde{w}) \leq \varepsilon$
   - $\widetilde{w'}$ is ε-close to w' *i.e.* $\mathrm{d}(w', \widetilde{w'}) \leq \varepsilon$

From $\mathscr{T}_1 \subseteq \mathscr{T}_2$ to $\mathscr{T}_1 \subseteq_\varepsilon \mathscr{T}_2$

1. **exact inclusion:** $(w, w') \in \mathscr{T}_1 \implies (w, w') \in \mathscr{T}_2$
2. **ε-inclusion:** $(w, w') \in \mathscr{T}_1 \implies \exists\, (\widetilde{w}, \widetilde{w'}) \in \mathscr{T}_2$ such that, for the classical edit distance with moves
   - $\widetilde{w}$ is ε-close to w *i.e.* $\mathrm{d}(w, \widetilde{w}) \leq \varepsilon$
   - $\widetilde{w'}$ is ε-close to w' *i.e.* $\mathrm{d}(w', \widetilde{w'}) \leq \varepsilon$

Finer ? Coarser ?

exact inclusion $\implies$ ε-inclusion

Equivalence is undecidable

1. The equivalence problem for non-deterministic transducers (one way, finite state) is undecidable even when they cannot read or write the empty string

   T. V. Griffiths, “The unsolvability of the Equivalence Problem for Lambda-Free nondeterministic generalized machines”,1968.

2. The equivalence problem for 1-free[2] input deterministic[3] sequential transducers is undecidable

   J. Karhumäki and L. P. Lisovik, “On the Equivalence of Finite Substitutions and Transducers”,1999, (Corollary 1).

[2] Transitions do not output the empty string

[3] The (input) underlying automaton is deterministic

Decidability

1. The equivalence for deterministic transducers is decidable
   Bird 1973, Vaillant 1974
2. The equivalence for n-tape deterministic finite automata is decidable Harju & Karhumäki, 1991
3. The equivalence for sequential transducers satisfying the prefix condition[4] is decidable Karhumäki & Lisovik, 1999
4. The equivalence on an NR set of strings for two deterministic two-way finite state transducers is decidable
   Engelfriet, Maneth, 2005

[4] Avoid two transitions from the same state, the same input letter, and an output different and prefix of the other

PSPACE-hardness

Equivalence is at least PSPACE-hard, even restricted to identity edges

proof

as hard as the equivalence of non-deterministic finite automata.

PSPACE-completeness for a weak form of non-determinism

The equivalence of deterministic two-way (sequential) transducers which are allowed to make some finite number of non-deterministic moves is a PSPACE-complete problem

pointed out by Gurari, 1982

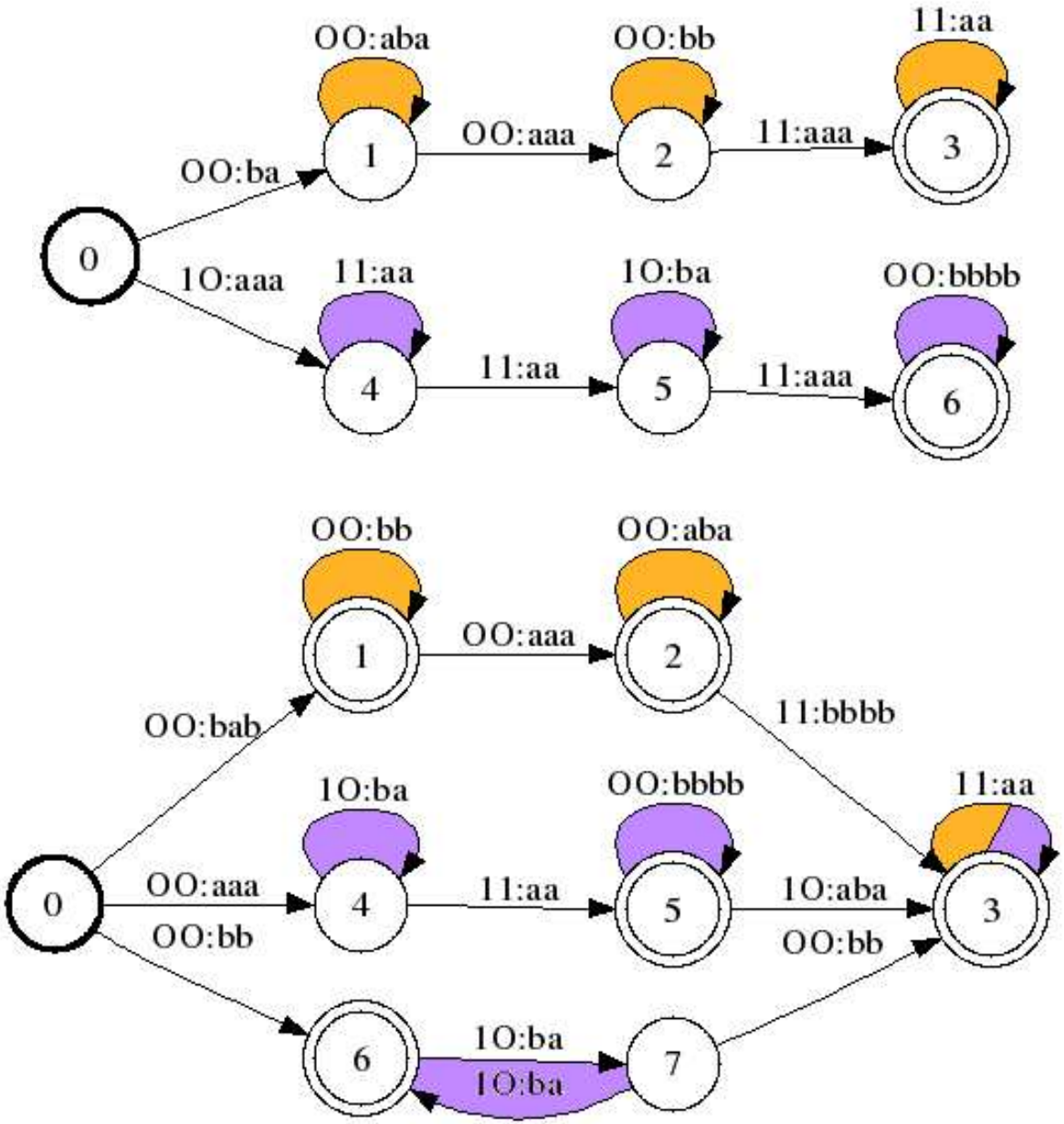
4 warm up

5 definition

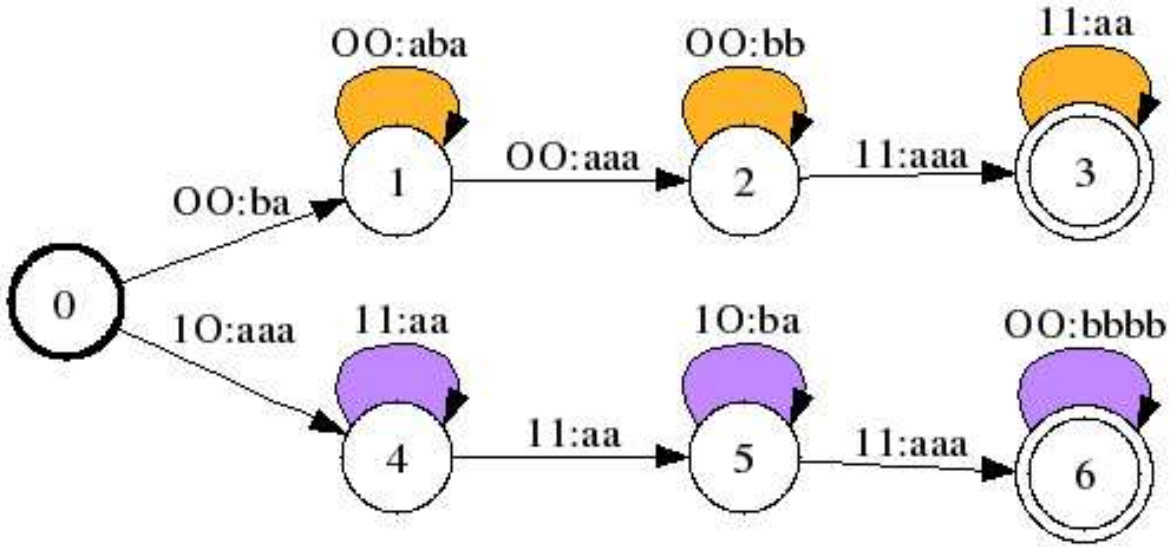
6 exact case

7 ε-equivalence

- two transducers ε-close
- statistical representation
- In the relation => close to the embedding
- In the embedding => close to the relation
- counterexamples : sources of non ε-equivalence


OO:aba
OO:bb
11:aa
OO:ba
OO:aaa
11:aaa
0
1
2
3
1O:aaa
11:aa
1O:ba
OO:bbbb
4
11:aa
5
11:aaa
6
OO:bb
OO:aba
1
OO:aaa
2
OO:bab
11:bbbb
1O:ba
OO:bbbb
11:aa
0
OO:aaa
4
11:aa
5
1O:aba
3
OO:bb
OO:bb
6
1O:ba
7
1O:ba

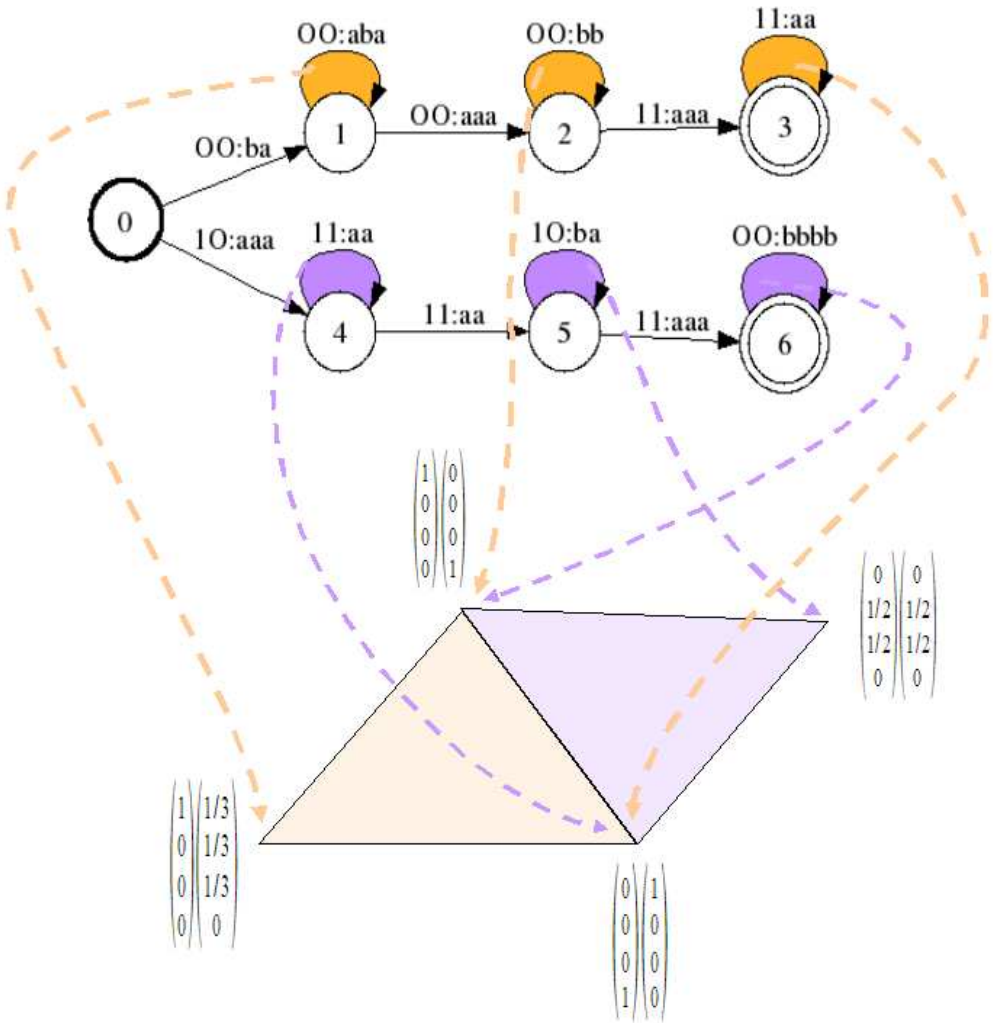
Forget finite contributions



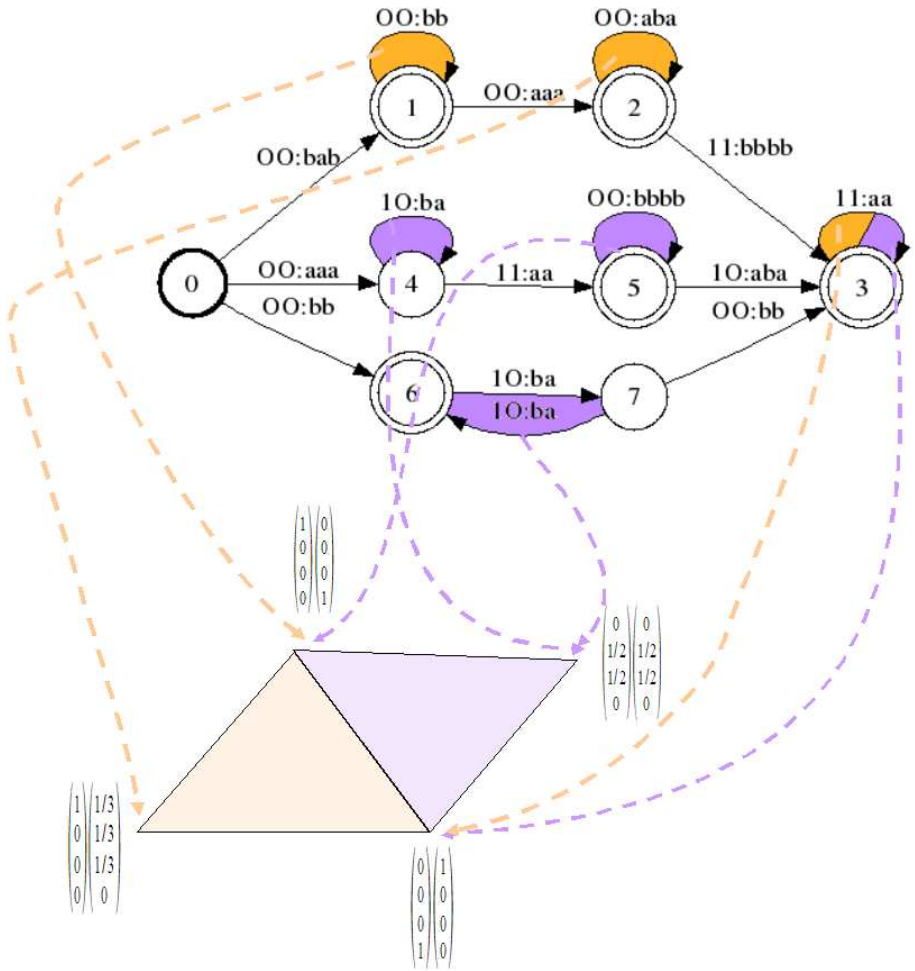
In essence:

$$t = (00 : aba)^*.(00 : bb)^*.(11 : aa)^* + (11 : aa)^*.(10 : ba)^*.(00 : bbbb)^*$$

$\mathscr{T}_1$

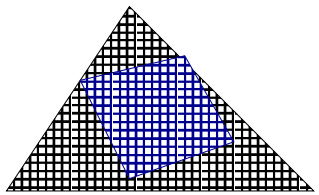

OO:aba
OO:bb
11:aa
OO:aaa
11:aaa
OO:ba
1
2
3
0
1O:aaa
11:aa
1O:ba
OO:bbbb
4
11:aa
5
11:aaa
6

$\mathscr{T}_2$


OO:bb
OO:aba
1
OO:aaa
2
OO:bab
11:bbbb
1O:ba
OO:bbbb
11:aa
0
OO:aaa
4
11:aa
5
1O:aba
3
OO:bb
OO:bb
6
1O:ba
7
1O:ba

ε-Inclusion

- Apply the k-powering of $\mathscr{T}_1$ and $\mathscr{T}_2$, $k = 1/\varepsilon$:
  - Matrix multiplication
- compute the embeddings $\mathcal{H}_\varepsilon = (H_\varepsilon^I, H_\varepsilon^O)$ in the statistical space:
  - analyze paths π in $\widehat{\mathscr{T}_i^k}$
  - convex hulls for π-compatible loops
- $H_\varepsilon^1 \subseteq H_\varepsilon^2$?:
  - H_ε^i is a union of polytopes
  - inclusion of the grids of step ε



- check lengths

$\Longrightarrow$ PSPACE tester for every fixed $\varepsilon > 0$

Assume $(w, w') \in \mathscr{T}^k$

- let π be a path in $\widehat{\mathscr{T}^k}$ s.t. $w' \in \mathscr{T}^k_\pi(w)$
- following π, $w = v_0 u_1 u_2 \dots u_l v_l$,
  $|v_0|, |u_1|, \dots |u_l|, |v_l| \leq m.k$, $\{u_1, \dots, u_l\}$: π-compatible
- $\{u_1, \dots, u_l\} = \{u'_1, \dots, u'_b\}$ with $b \leq a^k$
- take $\widetilde{w} = v_0 (u'_1)^{k_1} v_1 (u'_2)^{k_2} \dots v_{b-1} (u'_b)^{k_b} v_b$ with
  $v_0 v_1 \dots v_b$: witness from compatibility,
  v_i: sequences of less than m edges.
- dist$(w, \widetilde{w}) \leq (b+1)m + \varepsilon n$
- take $\widetilde{w'} = out(v_0)(out(u'_1))^{k_1} out(v_1) \dots (out(u'_b))^{k_b} out(v_b)$
- $(\widetilde{w}, \widetilde{w'}) \in \mathscr{T}^k$
- dist$(w', \widetilde{w'}) \leq maxout(b+1)m + \varepsilon n$

$(\mathsf{ustat}(\widetilde{w}), \mathsf{ustat}(\widetilde{w'}))$ is close to the polytope

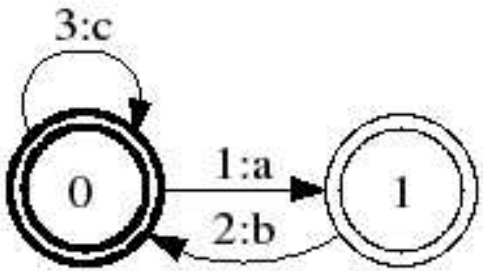
- $\widetilde{x} = \mathsf{ustat}(\widetilde{w})$, $\widetilde{x'} = \mathsf{ustat}(\widetilde{w'})$
- $\beta_i = \frac{k_i}{\sum_i k_i}$
- $\beta = (\beta_1, \ldots, \beta_b)$
- $x = Renorm(\sum_{i=1}^{b} \beta_i l_{in}(B_i) \mathcal{H}_{lim_{in}}(B_i))$
- $x \in \mathcal{H}^I$
- $|x - \widetilde{x}| \leq \frac{(b+1)(3(k-1)+m.p)}{\widetilde{n}}$
- $x' = Renorm(\sum_{i \in b} \frac{l_{out}(B_i)}{l_{in}(B_i)} \beta_i \mathcal{H}_{lim_{out}}(B_i))$
- $(x, x') \in \mathcal{H}$
- $|x' - \widetilde{x'}| \leq \frac{maxout((b+1)(3(k-1)+m.p))}{\widetilde{n'}}$

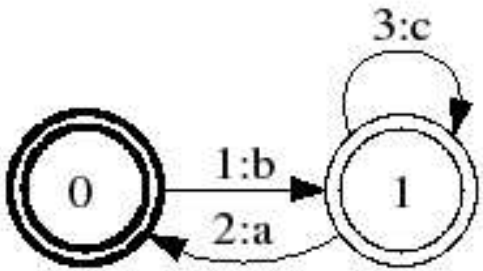
Assume $(\lambda, \lambda') \in \mathcal{H}$

- take sets of summits T, (λ, λ', π)-compatibles:
  - $\lambda = \sum_{i \in T} \lambda_i \mathcal{H}_{lim_{in}}(B_i)$
  - $\lambda' = \sum_{i \in T} \lambda'_i \mathcal{H}_{lim_{out}}(B_i)$
  - $\lambda' = Renorm\left(\frac{l_{out}(B_i)}{l_{in}(B_i)}\lambda_i\right)$
- approximate λ_i by rationals α_i
- denormalize $\widetilde{\lambda_i} = c.\frac{l_{out}(B_i)}{l_{in}(B_i)}\alpha_i$
- take $w = u_0(in(B_1))^{\widetilde{\lambda_1}} u_1 (in(B_2))^{\widetilde{\lambda_2}} u_2 \dots u_{|T|}$
- take $w' = u'_0(out(B_1))^{\widetilde{\lambda_1}} u'_1 (out(B_2))^{\widetilde{\lambda_2}} u'_2 \dots u'_{|T|}$
- $(w, w') \in \mathscr{T}$
- $|(\mathsf{ustat}(w), \mathsf{ustat}(w')) - (\lambda, \lambda')|_1 \leq 2(maxout + 1)\varepsilon$

Sources of non ε-equivalence

1. statistical domains ε-far

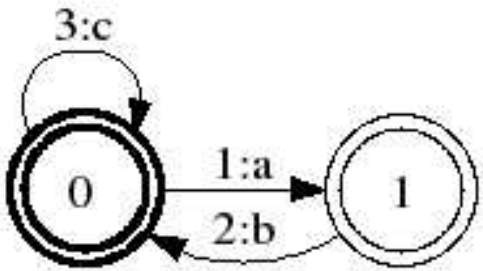


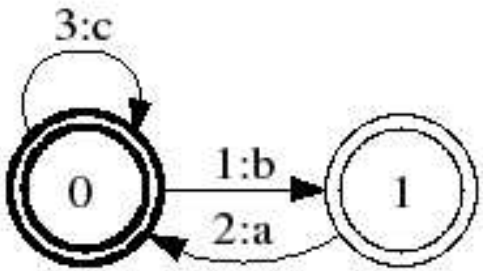


2. statistical images ε-far
3. generated length asymptotically different

Sources of non ε-equivalence

1. statistical domains ε-far
2. statistical images ε-far

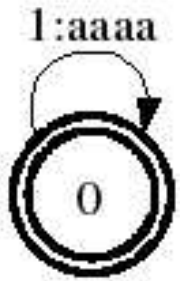




3. generated length asymptotically different

Sources of non ε-equivalence

1. statistical domains ε-far
2. statistical images ε-far
3. generated length asymptotically different



Sources of non ε-equivalence

1. statistical domains ε-far
2. statistical images ε-far
3. generated length asymptotically different

Solutions :

1. the domain is regular: similar embedding of a regular language: PTIME tester.
2. use the embedding
3. generating functions (rational functions)

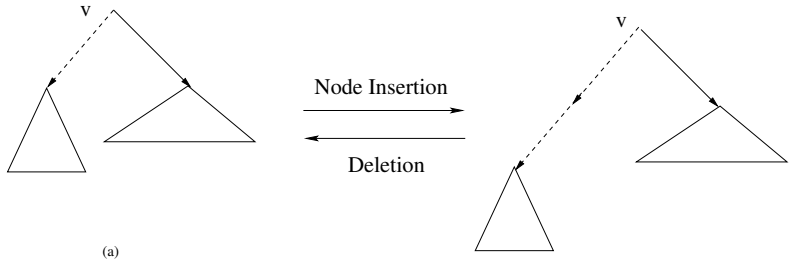
Outline: Extention to trees

8 Distances and Statistics
- Tree Edit Distance with Moves

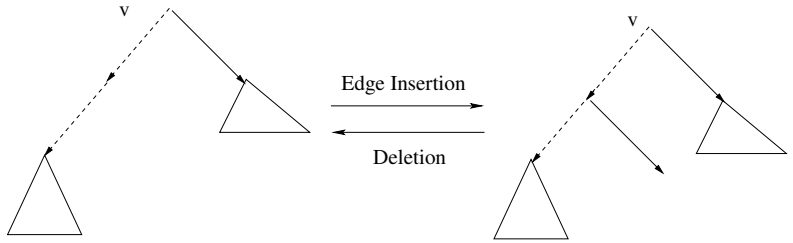
9 Tree Transducers
- Definition
- Example
- the next step

Tree Edit Distance with Moves : Elementary Operations

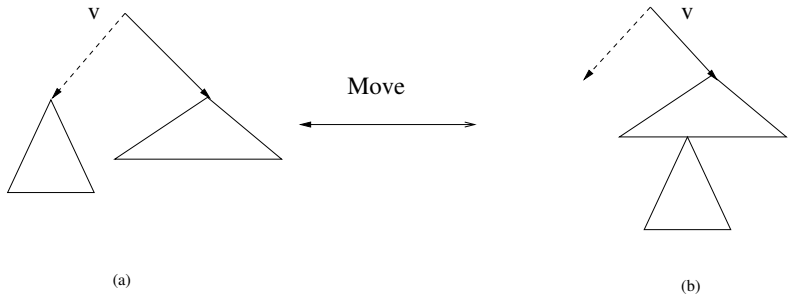
Node operation:



Edge operation:



Move:



Relaleling

Let $H_{\Sigma'}[Q]$ be the set of finite sequences of finite trees where one leaf of each tree is a distinguished element labeled by a sequence of states in Q, which is possibly empty.

Tree Transducers

A tree transducter $\mathscr{T}$ is a tuple $(\Sigma, \Sigma', Q, I, \delta)$ where

- Σ in the input alphabet, Σ' the output alphabet
- Q a set of states, $I \in Q$ an initial state
- $\delta : \Sigma \times Q \to H_{\Sigma'}[Q]$,

An xlst program

```
<xsl:output method="xml" indent="yes"/>

<xsl:template match="/">
    <root>  <s> <summary/>
                <xsl:apply-templates mode="q0"/>
            </s>
            <c> <xsl:apply-templates mode="q1"/>
            </c>
    </root>
    </xsl:template>

  </xsl:template match="a" mode="q0">
    <a/>
  </xsl:template>

  </xsl:template match="b" mode="q0">
    <b/>
  </xsl:template>

  </xsl:template match="a" mode="q1">
    <a>
       <xsl:apply-templates mode="q1"/>
    </a>
  </xsl:template>

  </xsl:template match="b" mode="q1">
    <b>
       <xsl:apply-templates mode="q1"/>
    </b>
  </xsl:template>
```

1. define tree statistics as matrices
2. Matrices distances & Edit distance
3. applications for xslt programs

thank you

$\mathcal{Q}$uestions ?

vieille@lri.fr