felipe@ rali Dernière modification January 25, 2015

Introduction à l'apprentissage analogique

Philippe Langlais

felipe@iro.umontreal.ca

Plan

Apprentissage Analogique

- Analogie
- Principe

Quelques réalisations

Sous le capot

- Définitions de l'analogie (formelle)
- Solveurs d'équations
- Recherche des analogies
- Filtrer le bruit

Applications

- Traduction de mots inconnus
- Traduction de termes
- Translitération

Apprentissage Analogique

- Analogie
- Principe

Quelques réalisations

Sous le capot

- Définitions de l'analogie (formelle)
- Solveurs d'équations
- Recherche des analogies
- Filtrer le bruit

Applications

- Traduction de mots inconnus
- Traduction de termes
- Translitération

Analogie

[x : y :: z : t]

“x est à y ce que z est à t”

Analogies sémantiques

[Christian : church :: Muslim : ?]

Analogies sémantiques

[Christian : church :: Muslim : mosque]

Analogies sémantiques

[Christian : church :: Muslim : mosque]

[3 : 4 :: 6 : ?]

Analogies sémantiques

[Christian : church :: Muslim : mosque]

[3 : 4 :: 6 : 8]

Analogies sémantiques

[Christian : church :: Muslim : mosque]

[3 : 4 :: 6 : 8]

[Paris : Vélib :: Montréal : ?]

Analogies sémantiques

[Christian : church :: Muslim : mosque]

[3 : 4 :: 6 : 8]

[Paris : Vélib :: Montréal : Bixi]

Analogies sémantiques

[Christian : church :: Muslim : mosque]

[3 : 4 :: 6 : 8]

[Paris : Vélib :: Montréal : Bixi]

Analogies sémantiques

[Christian : church :: Muslim : mosque]

[3 : 4 :: 6 : 8]

[Paris : Vélib :: Montréal : Bixi]

Analogies formelles

Exemples pris de (Lepage, 1998)

honor : honōrem :: ōrātor : ōrātōrem
reader : unreadable :: doer : undoable
lang : längste :: scharf : schärfste
répression : répressionnaire :: réaction : réactionnaire
croyons : créons :: mont royal : montréal
keras : mengeraskan :: kena : mengenakan

obj(big)& obj=circle : obj(small)⊂obj(big) & obj=circle :: obj(big)& obj=square : ?

? = obj(small)⊂obj(big) & obj=square

Apprentissage analogique

Principe d'après (Pirrelli & Yvon, 1999)

- un ensemble d'observations $\mathscr{L} = \{\langle s, t\rangle\}$
- une observation incomplete, $u = \langle s, ?\rangle$

Problem: Prédire les traits manquants de u

1. **Construire** $\mathscr{E}_{\mathscr{I}}(u)$**:**
   $\{(x, y, z) \in \mathscr{L}^3 \,|\, [\mathtt{I(x)} : \mathtt{I(y)} :: \mathtt{I(z)} : \mathtt{I(u)}]\}$
2. **Construire** $\mathscr{E}_{\mathscr{O}}(u)$**:**
   $\{t \,|\, [\mathtt{O(x)} : \mathtt{O(y)} :: \mathtt{O(z)} : \mathtt{t}], \forall (x, y, z) \in \mathscr{E}_{\mathscr{I}}(u)\}$
3. **Sélectionner les candidats de** $\mathscr{E}_{\mathscr{O}}(u)$.

where $I(\langle s, t\rangle) \equiv s$ and $O(\langle s, t\rangle) \equiv t$.

Illustration: parsing

$$\mathscr{L} \equiv \left\{ \begin{array}{l} (\texttt{S(Pr(elle).Vb(parle))}, \texttt{ she talks}), \\ (\texttt{S(Pr(elle).Vb(mange))}, \texttt{ she eats}), \\ (\texttt{S(Pr(elles).Vb(mangent))}, \texttt{ they eat}) \end{array} \right\}$$

Illustration: parsing

$$\mathscr{L} \equiv \left\{ \begin{array}{l} (\texttt{S(Pr(elle).Vb(parle))}, \texttt{ she talks}), \\ (\texttt{S(Pr(elle).Vb(mange))}, \texttt{ she eats}), \\ (\texttt{S(Pr(elles).Vb(mangent))}, \texttt{ they eat}) \end{array} \right\}$$

S(Pr(elles).Vb(parlent)) ?

Illustration: parsing

$$\mathscr{L} \equiv \left\{ \begin{array}{l} (\texttt{S(Pr(elle).Vb(parle))}, \texttt{ she talks}), \\ (\texttt{S(Pr(elle).Vb(mange))}, \texttt{ she eats}), \\ (\texttt{S(Pr(elles).Vb(mangent))}, \texttt{ they eat}) \end{array} \right\}$$

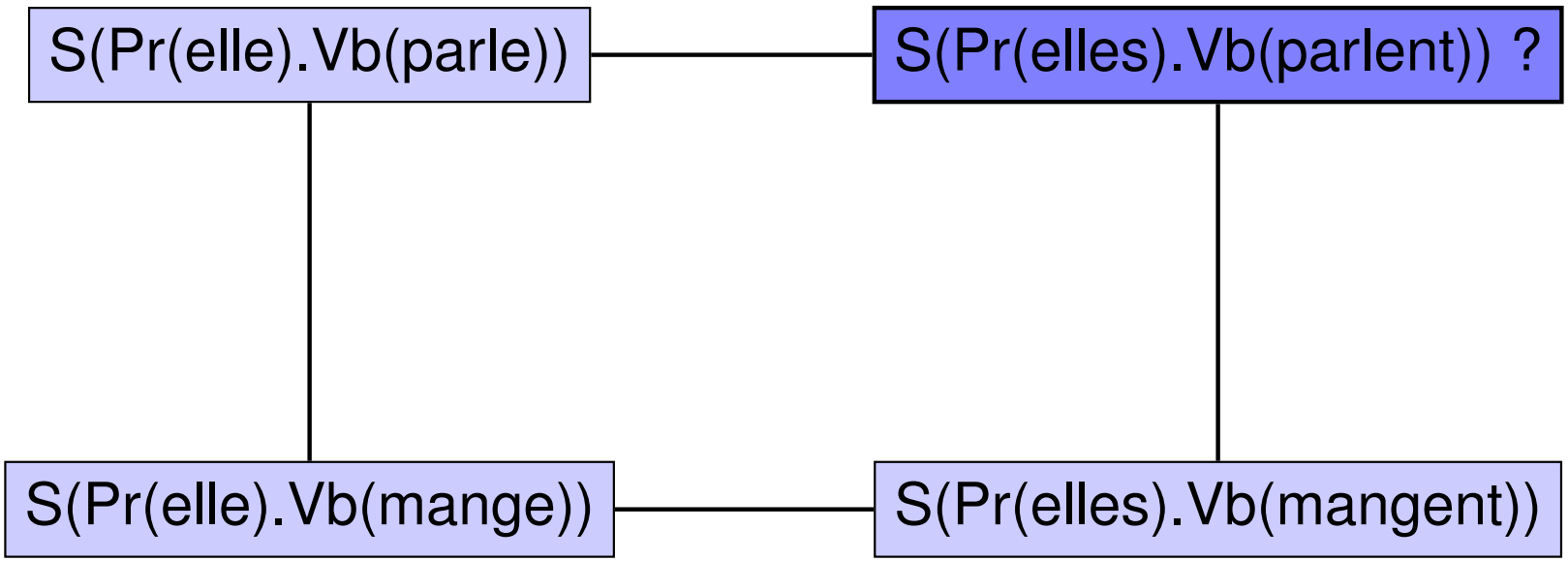


Illustration: parsing

$$\mathscr{L} \equiv \left\{ \begin{array}{l} (\texttt{S(Pr(elle).Vb(parle))}, \texttt{ she talks}), \\ (\texttt{S(Pr(elle).Vb(mange))}, \texttt{ she eats}), \\ (\texttt{S(Pr(elles).Vb(mangent))}, \texttt{ they eat}) \end{array} \right\}$$

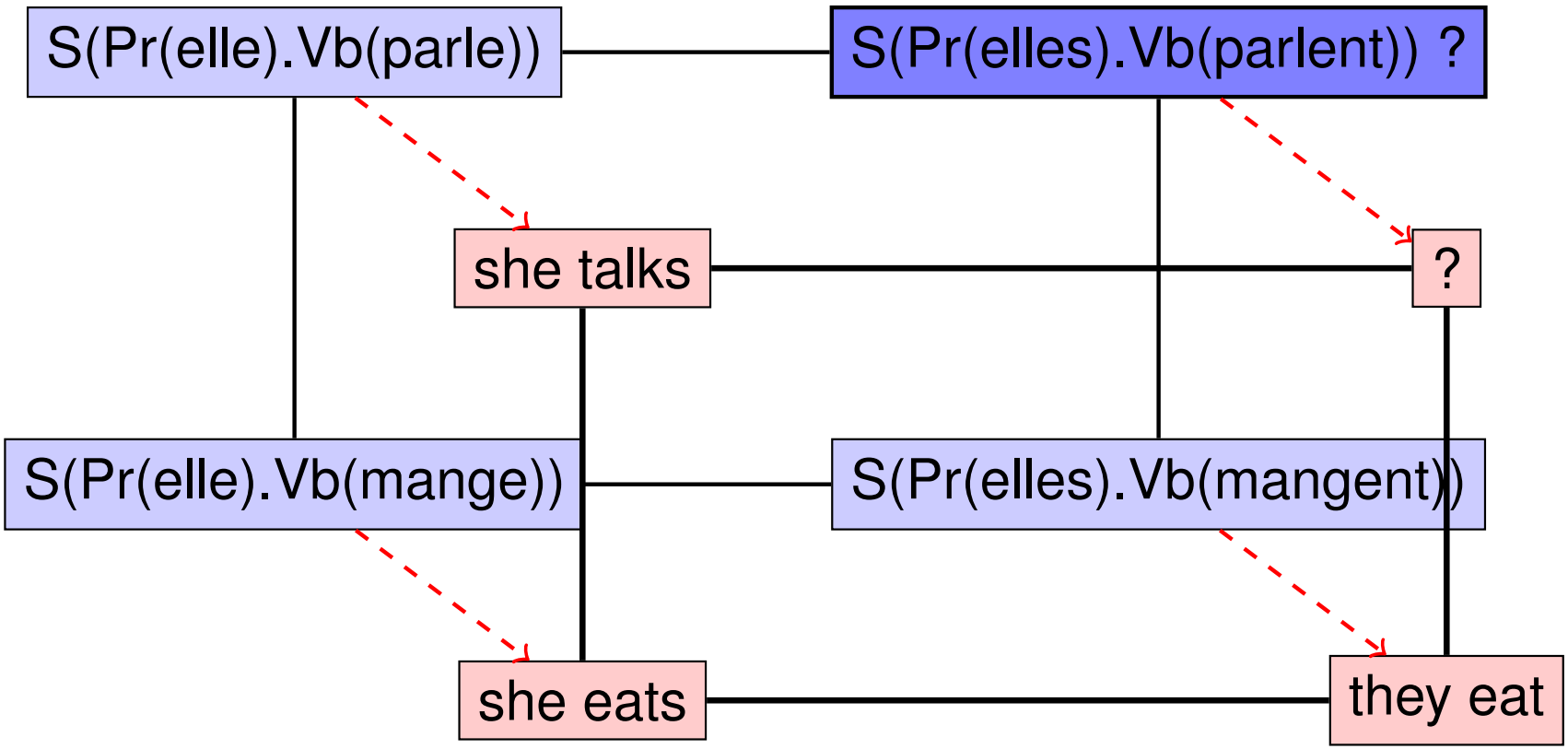


Illustration: parsing

$$\mathscr{L} \equiv \left\{ \begin{array}{l} (\texttt{S(Pr(elle).Vb(parle))}, \texttt{ she talks}), \\ (\texttt{S(Pr(elle).Vb(mange))}, \texttt{ she eats}), \\ (\texttt{S(Pr(elles).Vb(mangent))}, \texttt{ they eat}) \end{array} \right\}$$

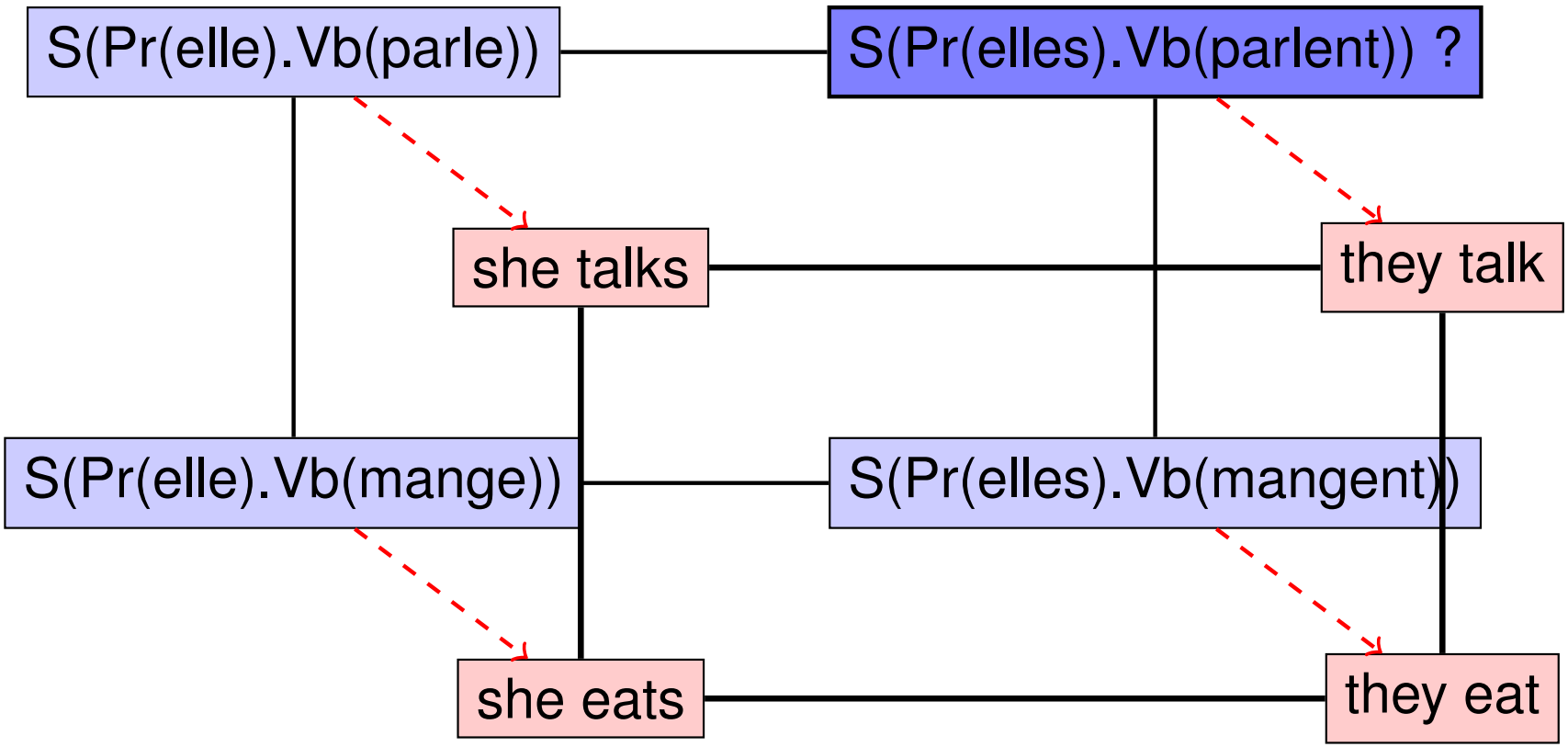


Illustration: traduction de mots inconnus

Step-1: calcul de $\mathscr{E}_{\mathscr{J}}$(futilité)

futilité?

Illustration: traduction de mots inconnus

Step-1: calcul de $\mathscr{E}_{\mathscr{J}}$(futilité)

faillites

faillite

futilité?

futilités

[faillites : faillite :: futilités : futilité]

Illustration: traduction de mots inconnus

Step-1: calcul de $\mathscr{E}_{\mathscr{J}}$(futilité)

faillites — futile

faillite

futilité?

futilités

lucide — lucidité

[lucide : lucidité :: futile : futilité]

Illustration: traduction de mots inconnus

Step-1: calcul de $\mathscr{E}_{\mathscr{J}}$(futilité)

faillites

futile

faillite

hostilités

futilité?

hostiles

futilités

lucide

lucidité

[hostiles : hostilités :: futile : futilité]

Illustration: traduction de mots inconnus

Step-1: calcul de $\mathscr{E}_{\mathscr{J}}$(futilité)

faillites
réalités
faillite
futilités
futilités
lucide

futilité?

futile
réalité
hostilités
hostiles
lucidité

[réalités : réalité :: futilités : futilité]

Illustration: traduction de mots inconnus

Step-1: calcul de $\mathscr{E}_{\mathscr{J}}$(futilité)

faillites brutale futile
réalités brutalité réalité mutilé
faillite bestialité hostilités
futilités **futilité?** natalité
facile facilités hostiles
futilités bestiale timide natale
lucide faillite maille
timidité lucidité

2944 équations considérées, 84 valides

Illustration: traduction de mots inconnus

Step-1: calcul de $\mathscr{E}_{\mathscr{J}}$ (futilité)

⟨activités, activité, futilités⟩ ⟨fatale, futile, fatalité⟩ ⟨timide, timidité, futile⟩

⟨faciles, facilités, futile⟩ ⟨cupide, cupidité, futile⟩ ⟨utilisés, futilités, utilisé⟩

⟨humide, humidité, futile⟩ ⟨brutalités, brutalité, futilités⟩ ⟨multipliés, multiplié, futilités⟩

⟨qualités, qualité, futilités⟩ ⟨totale, totalité, futile⟩ ⟨utilisés, utilisé, futilités⟩

⟨mutilés, mutilé, futilités⟩ ⟨félicités, futilités, félicité⟩ ⟨active, activité, futile⟩

⟨mature, maturité, futile⟩ ⟨unités, futilités, unité⟩ ⟨habilités, habilité, futilités⟩

⟨fragile, fragilité, futile⟩ ⟨subtile, subtilité, futile⟩ ⟨mutilés, futilités, mutilé⟩

⟨autorités, autorité, futilités⟩ ⟨vitale, vitalité, futile⟩ ⟨autorisés, autorisé, futilités⟩

⟨maille, faillite, mutilé⟩ ⟨mute, mutilé, futile⟩ ⟨facultés, faculté, futilités⟩

⟨félicités, félicité, futilités⟩ ⟨utile, futile, utilité⟩ ⟨facilités, facilité, futilités⟩

⟨rurale, ruralité, futile⟩ ⟨brutalités, futilités, brutalité⟩ ⟨finales, finalités, futile⟩

⟨subtiles, subtilités, futile⟩ ⟨spatiale, spatialité, futile⟩ ⟨visités, visité, futilités⟩

⟨réalités, réalité, futilités⟩ ⟨pénale, pénalité, futile⟩ ⟨brutales, brutalités, futile⟩

⟨habilités, futilités, habilité⟩ ⟨hostilités, futilités, hostilité⟩ (+ 42 autres . . .)

Illustration: traduction de mots inconnus

Step-2: calcul de $\mathscr{E}_O$(futilité)

[faillites : faillite :: futilités : futilité]

faillites	$\leftrightarrow$	**bankruptcies**, bankruptcy
faillite	$\leftrightarrow$	collapse, bust, insolvency, ruin, complaining, bankrupt, business, **bankruptcy**, wall
futilités	$\leftrightarrow$	**trivialities**

[bankruptcies : bankruptcy :: trivialities : ?] { **triviality** }

Illustration: traduction de mots inconnus

Step-2: calcul de E_O(futilité)

[lucide : lucidité :: futile : futilité]

lucide	$\leftrightarrow$	clear-thinking, blind, refocussed, lucid, far-sighted
lucidité	$\leftrightarrow$	well-informed, meticulously, moderation, penetrating, lucidity, clear-headed, clear-sighted
futilité	$\leftrightarrow$	meaningless, futile

[lucid : lucidity :: meaningless : ?] { meaninglessity }
[lucid : lucidity :: futile : ?] { futiityle, futileity, futilitye }
[far-sighted : clear-sighted :: futile : ?] {fucleile, cleutile, clefuile, cluteile, culetile, cluetile, cfuleile, . . . }

Illustration: traduction de mots inconnus

Step-3: sélection depuis $\mathscr{E}_O$(futilité)

1124 $\neq$ formes générées, 120 au moins 2 fois

(trivialities,27)
(triviality,14)
(futile,9)
(futilitye,9)
(meaningless,9)
(meaninglityess,8)
(trivialitie,6)
(futility,4)
(meaninglityss,4)
(futilities,3)
(meaninglesit,3)
(meaninglitiesss,3)
(meaningiltyss,2)
(futiilty,2)
(applicatitrivialitis,2)
(futiitiles,2)
(futiitly,2)
(futiityl,2)
(futileity,2)
(futilessne,2)
(high-triviality,2)
(ltbbyirivialitig,2)
(meaniitnglyss,2)
(superfluous,2)
(meaningitilesss,2)
(meaningitlyss,2)
(meaninglesity,2)
(meaninglessity,2)
(meaninglityes,2)
(meaninigltyss,2)
(meaninitglyss,2)
(meaninitiglesss,2)
(mfaninglecilits,2)

...

Illustration: traduction de mots inconnus

Step-3: sélection depuis $\mathscr{E}_O$(futilité)

1124 $\neq$ formes générées, 120 au moins 2 fois

(trivialities,27)
(triviality,14)
(futile,9)
~~(futilitye,9)~~
(meaningless,9)
~~(meaninglityess,8)~~
(trivialitie,6)
(futility,4)
~~(meaninglityss,4)~~
(futilities,3)
~~(meaninglesit,3)~~
~~(meaninglitiesss,3)~~
~~(meaningiltyss,2)~~
~~(futiilty,2)~~
~~(applicatitrivialitis,2)~~
~~(futiitiles,2)~~
~~(futiitly,2)~~
~~(futiityl,2)~~
~~(futileity,2)~~
~~(futilessne,2)~~
~~(high-triviality,2)~~
~~(ltbbyirivialitig,2)~~
~~(meaniitnglyss,2)~~
(superfluous,2)
~~(meaningitilesss,2)~~
~~(meaningitlyss,2)~~
~~(meaninglesity,2)~~
~~(meaninglessity,2)~~
~~(meaninglityes,2)~~
~~(meaninigltyss,2)~~
~~(meaninitglyss,2)~~
~~(meaninitiglesss,2)~~
~~(mfaninglecilits,2)~~
...

Illustration: traduction de mots inconnus

Step-3: sélection depuis $\mathscr{E}_O$(futilité)

13 solutions conservées

(trivialities,27) (triviality,14) (futile,9)
(meaningless,9) (futility,4) (meaninglessness,3)
(superfluous,2) (unwieldy,2) (unnecessary,2)
(uselessness,2) (trivially,1) (tie,1) (trivial,1)

Quelques idées fausses

- analogie $\neq$ similarité ([x : y :: x : y])
- analogie identifiées dans chaque espace (source et cible) **indépendamment**

  biais inductif: homomorphisme
- ressemblances avec les k plus proches voisins, mais:
  - aucune distance
  - approche réversible
  - espace de sortie peut être structuré
- une analogie **fortuite** n'invalide pas l'apprentissage !

  [suddenly, she appeared : she appeared all of a sudden :: immediately, she appeared : she appeared all of a immediate]

Pourquoi est-ce séduisant en TALN?

- Un corpus est plus qu'une séquence de **mots**
- Vue paradigmatique des données:
  - [funny : funniest :: lucky : luckiest]
  - [suddenly, she appeared : she appeared suddenly :: today, she appeared : she appeared today]
  - [This guy drinks too much : This boat sank :: These guys drink too much : These boats sank]
  - [Elle est émue : Il est ému :: Elle est touchée : Il est touché]
- Aucune connaissance à utiliser pour définir la correspondance entre les deux espaces

Lire [Pirrelli & Yvon, 1999]

Apprentissage Analogique

- Analogie
- Principe

Quelques réalisations

Sous le capot

- Définitions de l'analogie (formelle)
- Solveurs d'équations
- Recherche des analogies
- Filtrer le bruit

Applications

- Traduction de mots inconnus
- Traduction de termes
- Translitération

Analogies sémantiques et tests SAT

(Turney & Littman, 2005; Turney, 2006)

stem	mason : stone
a)	teacher : chalk
b)	carpenter : wood
c)	soldier : gun
d)	photograph : camera
e)	book : word

http://www.sadlier-oxford.com/phonics/analogies/analogiesx.htm

Analogies sémantiques et tests SAT

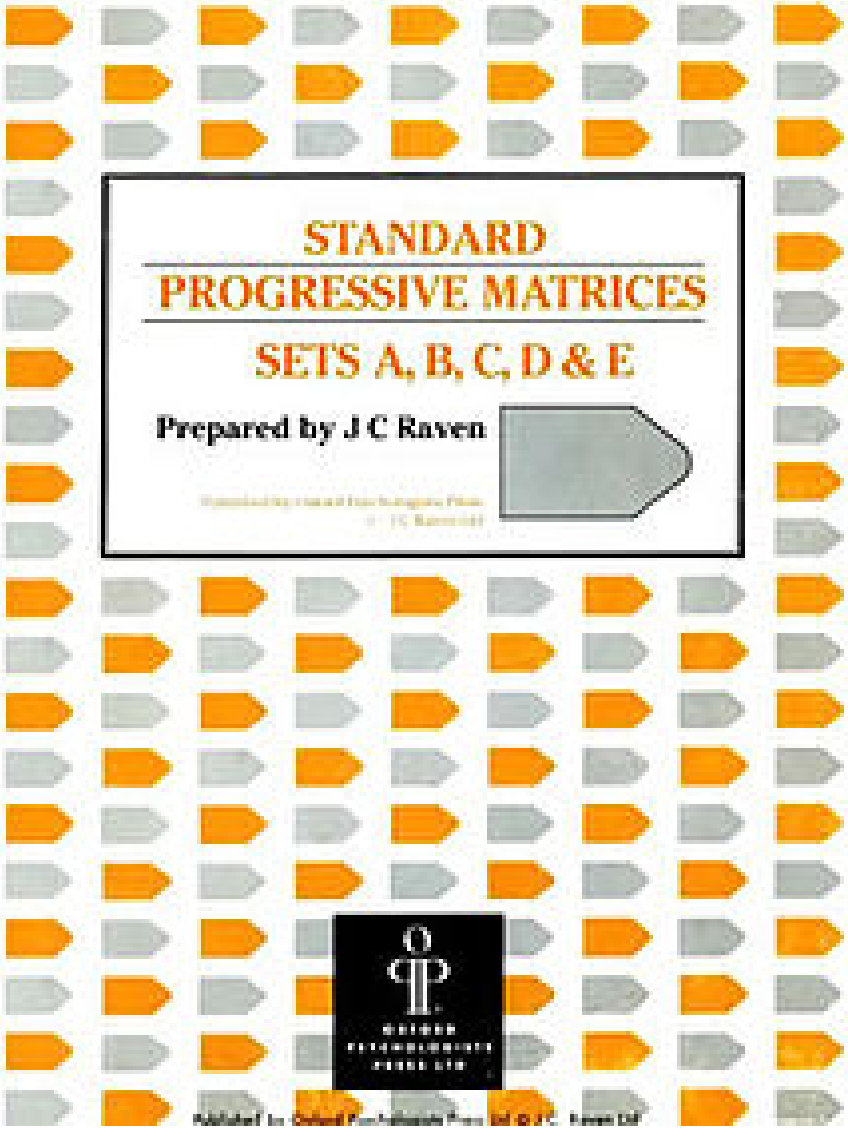
(Turney & Littman, 2005; Turney, 2006)

stem	mason : stone
a)	teacher : chalk
b)	carpenter : wood
c)	soldier : gun
d)	photograph : camera
e)	book : word

http://www.sadlier-oxford.com/phonics/analogies/analogiesx.htm

Tests de Raven

Voir les slides de Correa, Prade & Richard



ALEPH (Lepage & Denoual, 2005)

Src: *It floated across the river*

: :: : It floated across the river

ALEPH (Lepage & Denoual, 2005)

Src: *It floated across the river*

	:		::		:	
They swam in the sea	:	They swam across the river	::	It floated in the sea	:	It floated across the river

ALEPH (Lepage & Denoual, 2005)

Src: *It floated across the river*

They swam in the sea	:	They swam across the river	::	It floated in the sea	:	It floated across the river
↕		↕		↕		↕
Nadaron en el mar	:	Atraversaron el rio nadando	::	Flotó en el mar	:	x

ALEPH (Lepage & Denoual, 2005)

Src: *It floated across the river*

They swam in the sea	:	They swam across the river	::	It floated in the sea	:	It floated across the river
↕		↕		↕		↕
Nadaron en el mar	:	Atraversaron el rio nadando	::	Flotó en el mar	:	x

Tgt: Atraversó el rio flotando

Prolog est de retour !

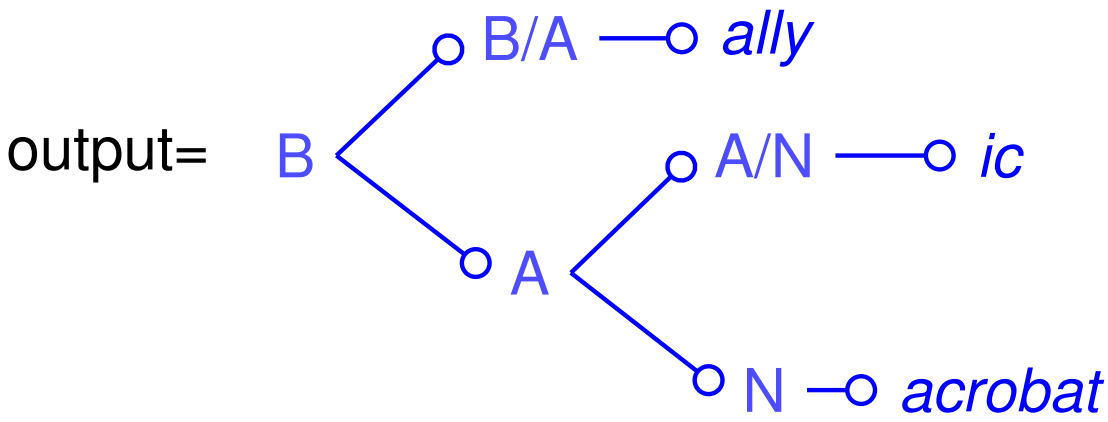
% bitext
translation(s_1,$\hat{s}_1$).
translation(s_2,$\hat{s}_2$).
...
translation(s_n,$\hat{s}_n$).

% ALEPH
translation(D,$\hat{D}$) :-
translation(A,$\hat{A}$)
translation(B,$\hat{B}$)
translation(C,$\hat{C}$)
analogie(A,B,C,D)
analogie($\hat{A}$,$\hat{B}$,$\hat{C}$,$\hat{D}$)
assert(translation(D,$\hat{D}$)).

(Stroppa & Yvon, 2005)

- **Task 1:** mot inconnu $\rightarrow$ traits morpho-syntaxiques
  input = replying
  output = {reply, V-PP}

- **Task 2:** mot inconnu $\rightarrow$ analyse morphologique
  input= acrobatically



(Moreau et al, 2007)

- $m_1 =$ désinstaller $\qquad m_2 =$ réinstallation
  $m_2 = m_1$ - dés$_p$ + ré$_p$ - er$_s$ + ation$_s$
- $m_1 =$ désydrater

(Moreau et al, 2007)

- $m_1 =$ désinstaller $\quad m_2 =$ réinstallation
  $m_2 = m_1$ - dés$_p$ + ré$_p$ - er$_s$ + ation$_s$
- $m_1 =$ désydrater $\quad m_2 =$ réhydratation

(Moreau et al, 2007)

- $m_1 =$ désinstalleR $\qquad m_2 =$ réinstallation
  $m_2 = m_1$ - dés$_p$ + ré$_p$ - er$_s$ + ation$_s$
- $m_1 =$ désydrater

ENTRAÎNEMENT

1. soit d un document de la collection C
2. soit les mots de d qui partagent une grande sous-chaîne ($\geq$ 7 car.)
3. mémoriser les correspondances
4. `aller en` 1

EXPANSION DE REQUÊTE

1. Soit m un mot de la requête r
2. Appliquer les règles à $m \Rightarrow S$
3. Ajouter à r les mots de $S \in$ voc(C)

- **Ex:** *pollution des eaux souterraines*

↪ *pollution des eaux souterraines polluants dépollution anti-pollution pollutions polluées polluent eau souterraine souterrains souterrain*

Gains consistants pour des collections de petite taille dans 6 langues (30 requètes)

Définitions de l'analogie formelle

(Pirrelli & Yvon, 1999)

- Symboles:

$$[\mathtt{x} : \mathtt{y} :: \mathtt{z} : \mathtt{t}] \iff \text{or} \begin{cases} x = y \text{ and } z = t \\ x = z \text{ and } y = t \end{cases}$$

[Noun : Noun :: Verb : Verb] , [Noun : Verb :: Noun : Verb]

- Chaînes:

$$[\mathtt{x} : \mathtt{y} :: \mathtt{z} : \mathtt{t}] \iff \text{or} \begin{cases} x = bc, y = bd, z = ac, t = ad \\ x = bc, y = ac, z = bd, t = ad \end{cases}$$

[dream : dreamer :: eat : eater] , [steal : ceal :: stage : cage]

- S'étend aux ensembles de valeurs

Définition de l'analogie formelle

(Lepage, 1998)

- chaînes:

$$[\mathtt{x} : \mathtt{y} :: \mathtt{z} : \mathtt{t}] \Rightarrow \left[\begin{array}{lcl} \sigma(y,t) & = & -|x| + |y| + \sigma(x,z) \\ \sigma(z,t) & = & -|x| + |z| + \sigma(x,y) \\ \sigma(x,y,z,t) & = & -|x| + \sigma(x,y) + \sigma(x,z) \\ |t|_a & = & -|x|_a + |y|_a + |z|_a \; \forall a \end{array}\right.$$

$[\mathtt{believer} : \mathtt{unbelievable} :: \mathtt{dreamer} : \mathtt{undreamable}]$

- Formalisé par un algorithme

Définition de l'analogie formelle

(Stroppa & Yvon, 2005)

Pour tout $(x, y, z, t) \in \Sigma^{\star 4}$, $[\mathtt{x} : \mathtt{y} :: \mathtt{z} : \mathtt{t}]$ **ssi** on peut trouver une factorisation $(f_x, f_y, f_z, f_t) \in (\Sigma^{\star d})^4$ telle que $\forall i \in [1, d]$:

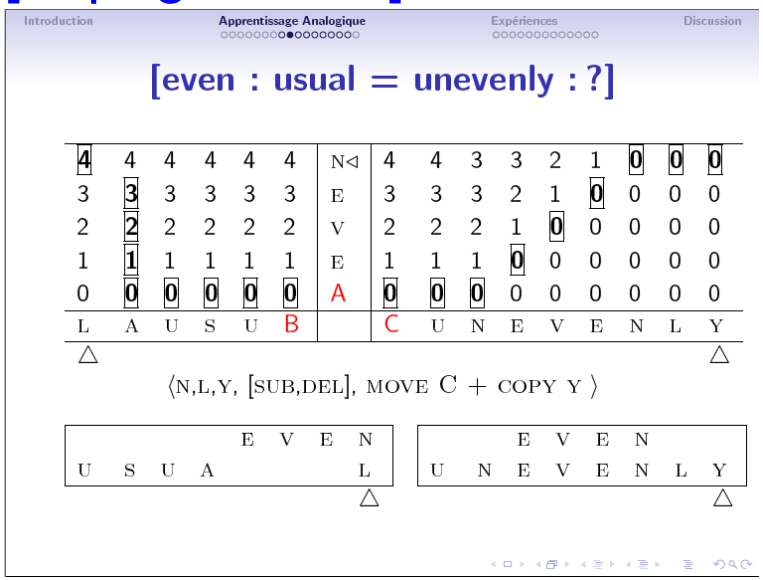
$$(f_y^{(i)}, f_z^{(i)}) \in \left\{ (f_x^{(i)}, f_t^{(i)}), (f_t^{(i)}, f_x^{(i)}) \right\}$$

- le plus petit d pour lequel cela est vrai est appelé le **degré**
- $f_x^{(i)}, f_y^{(i)}, f_z^{(i)}$ and $f_t^{(i)}$ sont appelés les facteurs

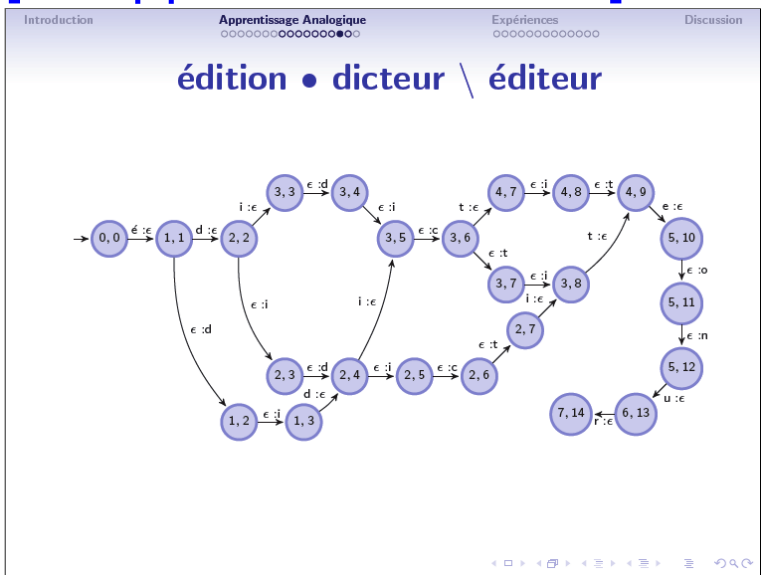
x $\equiv$	th	is	guy	ε	dr	inks	too much
y $\equiv$	th	is	boat	ε	s	inks	ε
z $\equiv$	th	ese	guy	s	dr	unk	too much
t $\equiv$	th	ese	boat	s	s	unk	ε

Deux solveurs génériques

[Lepage, 1998]



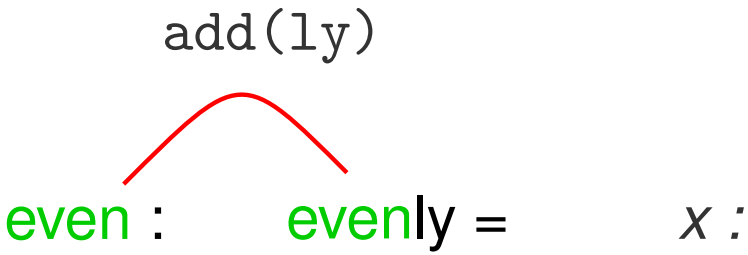
[Stroppa & Yvon, 2005]



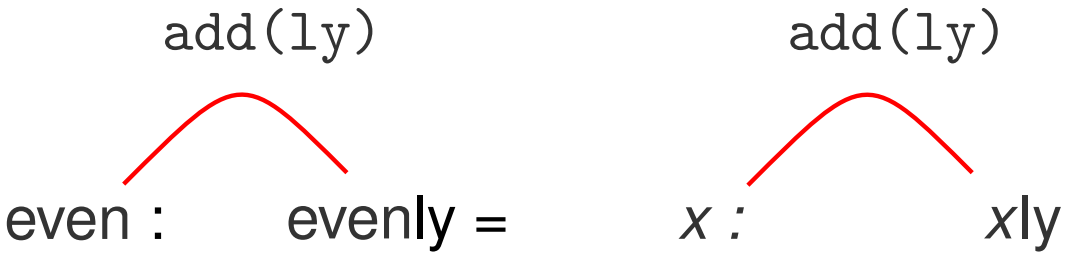
Intuition

even : evenly = *x :*

Intuition



Intuition

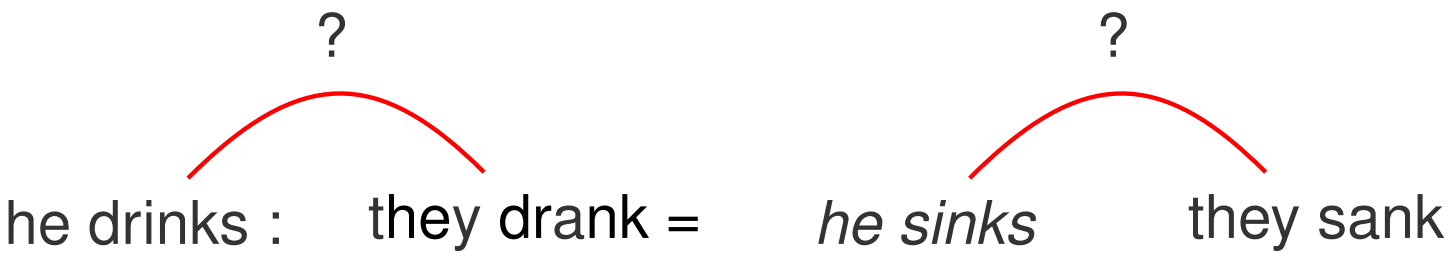


Intuition

he drinks : they drank = *he sinks*

Intuition

Intuition



[even : usual = unevenly : ?]

4	4	4	4	4	4	N◁	4	4	3	3	2	1	**0**	**0**	**0**
3	**3**	3	3	3	3	E	3	3	3	2	1	**0**	0	0	0
2	**2**	2	2	2	2	V	2	2	2	1	**0**	0	0	0	0
1	**1**	1	1	1	1	E	1	1	1	**0**	0	0	0	0	0
0	**0**	**0**	**0**	**0**	**0**	A	**0**	**0**	**0**	0	0	0	0	0	0
L	A	U	S	U	B		C	U	N	E	V	E	N	L	Y
△															△

⟨N,L,Y, [SUB,DEL], MOVE C + COPY Y ⟩

				E	V	E	N		
U	S	U	A				L		
							△		

				E	V	E	N		
U	N	E	V	E	N	L	Y		
							△		

[even : usual = unevenly : ?]

4	4	4	4	4	4	N◁	4	4	3	3	2	1	**0**	**0**	**0**
3	**3**	3	3	3	3	E	3	3	3	2	1	**0**	0	0	0
2	**2**	2	2	2	2	V	2	2	2	1	**0**	0	0	0	0
1	**1**	1	1	1	1	E	1	1	1	**0**	0	0	0	0	0
0	**0**	**0**	**0**	**0**	**0**	A	**0**	**0**	**0**	0	0	0	0	0	0
L	A	U	S	U	B		C	U	N	E	V	E	N	L	Y
△														△	

⟨N,L,L, [SUB,DEL], MOVE C + COPY L ⟩

				E	V	E	N		
U	S	U	A				L		
							△		

			E	V	E	N			
U	N	E	V	E	N	L	Y		
						△			

Y

[even : usual = unevenly : ?]

4	4	4	4	4	4	N◁	4	4	3	3	2	1	**0**	**0**	**0**
3	**3**	3	3	3	3	E	3	3	3	2	1	**0**	0	0	0
2	**2**	2	2	2	2	V	2	2	2	1	**0**	0	0	0	0
1	**1**	1	1	1	1	E	1	1	1	**0**	0	0	0	0	0
0	**0**	**0**	**0**	**0**	**0**	A	**0**	**0**	**0**	0	0	0	0	0	0
L	A	U	S	U	B		C	U	N	E	V	E	N	L	Y
△													△		

⟨N,L,N, [SUB,<=>], MOVE A,B,C + COPY L ⟩

				E	V	E	N		
U	S	U	A				L		
							△		

				E	V	E	N		
U	N	E	V	E	N	L	Y		
							△		

LY

[even : usual = unevenly : ?]

4	4	4	4	4	4	N	4	4	3	3	2	1	**0**	**0**	**0**
3	**3**	3	3	3	3	E◁	3	3	3	2	1	**0**	0	0	0
2	**2**	2	2	2	2	V	2	2	2	1	**0**	0	0	0	0
1	**1**	1	1	1	1	E	1	1	1	**0**	0	0	0	0	0
0	**0**	**0**	**0**	**0**	**0**	A	**0**	**0**	**0**	0	0	0	0	0	0
L	A	U	S	U	B		C	U	N	E	V	E	N	L	Y
	△											△			

⟨E,A,E, [INS,<=>], MOVE A,C ⟩

						E	V	E	N				E	V	E	N		
U	S	U	A						L	U	N	E	V	E	N	L	Y
								△							△		

LLY

[even : usual = unevenly : ?]

4	4	4	4	4	4	N	4	4	3	3	2	1	**0**	**0**	**0**
3	**3**	3	3	3	3	E	3	3	3	2	1	**0**	0	0	0
2	**2**	2	2	2	2	V◁	2	2	2	1	**0**	0	0	0	0
1	**1**	1	1	1	1	E	1	1	1	**0**	0	0	0	0	0
0	**0**	**0**	**0**	**0**	**0**	A	**0**	**0**	**0**	0	0	0	0	0	0
L	A	U	S	U	B		C	U	N	E	V	E	N	L	Y
	△										△				

⟨V,A,V, [INS,<=>], MOVE A,C ⟩

						E	V	E	N						
U	S	U	A						L	U	N	E	V	E	N	L	Y
					△								△		

LLY

[even : usual = unevenly : ?]

4	4	4	4	4	4	N	4	4	3	3	2	1	**0**	**0**	**0**
3	**3**	3	3	3	3	E	3	3	3	2	1	**0**	0	0	0
2	**2**	2	2	2	2	V	2	2	2	1	**0**	0	0	0	0
1	**1**	1	1	1	1	E◁	1	1	1	**0**	0	0	0	0	0
0	**0**	**0**	**0**	**0**	**0**	A	**0**	**0**	**0**	0	0	0	0	0	0
L	A	U	S	U	B		C	U	N	E	V	E	N	L	Y
	△									△					

⟨E,A,E, [INS,<=>], MOVE A,C ⟩

					E	V	E	N				E	V	E	N
U	S	U	A					L	U	N	E	V	E	N	L	Y
			△								△				

LLY

[even : usual = unevenly : ?]

4	4	4	4	4	4	N	4	4	3	3	2	1	**0**	**0**	**0**
3	**3**	3	3	3	3	E	3	3	3	2	1	**0**	0	0	0
2	**2**	2	2	2	2	V	2	2	2	1	**0**	0	0	0	0
1	**1**	1	1	1	1	E	1	1	1	**0**	0	0	0	0	0
0	**0**	**0**	**0**	**0**	**0**	A◁	**0**	**0**	**0**	0	0	0	0	0	0
L	A	U	S	U	B		C	U	N	E	V	E	N	L	Y

△ △

⟨ε,A,N, [DEL,DEL], MOVE C + COPY N ⟩

				E	V	E	N		
U	S	U	A				L		

△

		E	V	E	N				
U	N	E	V	E	N	L	Y		

△

LLY

[even : usual = unevenly : ?]

4	4	4	4	4	4	N	4	4	3	3	2	1	**0**	**0**	**0**
3	**3**	3	3	3	3	E	3	3	3	2	1	**0**	0	0	0
2	**2**	2	2	2	2	V	2	2	2	1	**0**	0	0	0	0
1	**1**	1	1	1	1	E	1	1	1	**0**	0	0	0	0	0
0	**0**	**0**	**0**	**0**	**0**	A◁	**0**	**0**	**0**	0	0	0	0	0	0
L	A	U	S	U	B		C	U	N	E	V	E	N	L	Y

△ △

⟨ε,A,U, [DEL,DEL], MOVE C + COPY U ⟩

				E	V	E	N		
U	S	U	A				L	U	N

△

		E	V	E	N		
U	N	E	V	E	N	L	Y

△

NLLY

[even : usual = unevenly : ?]

4	4	4	4	4	4	N	4	4	3	3	2	1	**0**	**0**	**0**
3	**3**	3	3	3	3	E	3	3	3	2	1	**0**	0	0	0
2	**2**	2	2	2	2	V	2	2	2	1	**0**	0	0	0	0
1	**1**	1	1	1	1	E	1	1	1	**0**	0	0	0	0	0
0	**0**	**0**	**0**	**0**	**0**	A◁	**0**	**0**	**0**	0	0	0	0	0	0
L	A	U	S	U	B		C	U	N	E	V	E	N	L	Y

△ △

$\langle \varepsilon, A, \varepsilon, [\text{DEL}, \bullet], \text{MOVE B + COPY A} \rangle$

				E	V	E	N		
U	S	U	A				L		

△

			E	V	E	N			
U	N	E	V	E	N	L	Y		

△

UNLLY

[even : usual = unevenly : ?]

4	4	4	4	4	4	N	4	4	3	3	2	1	**0**	**0**	**0**
3	**3**	3	3	3	3	E	3	3	3	2	1	**0**	0	0	0
2	**2**	2	2	2	2	V	2	2	2	1	**0**	0	0	0	0
1	**1**	1	1	1	1	E	1	1	1	**0**	0	0	0	0	0
0	**0**	**0**	**0**	**0**	**0**	A◁	**0**	**0**	**0**	0	0	0	0	0	0
L	A	U	S	U	B		C	U	N	E	V	E	N	L	Y

△ △

$\langle \varepsilon, \text{U}, \varepsilon, [\text{DEL}, \bullet], \text{MOVE B + COPY U} \rangle$

				E	V	E	N		
U	S	U	A				L		

△

			E	V	E	N			
U	N	E	V	E	N	L	Y		

△

AUNLLY

[even : usual = unevenly : ?]

4	4	4	4	4	4	N	4	4	3	3	2	1	**0**	**0**	**0**
3	**3**	3	3	3	3	E	3	3	3	2	1	**0**	0	0	0
2	**2**	2	2	2	2	V	2	2	2	1	**0**	0	0	0	0
1	**1**	1	1	1	1	E	1	1	1	**0**	0	0	0	0	0
0	**0**	**0**	**0**	**0**	**0**	A◁	**0**	**0**	**0**	0	0	0	0	0	0
L	A	U	S	U	B		C	U	N	E	V	E	N	L	Y

△ △

$\langle \varepsilon, S, \varepsilon, [\text{DEL}, \bullet], \text{MOVE B +COPY S} \rangle$

						E	V	E	N						
U	S	U	A						L	U	N	E	V	E	N	L	Y
△										△							

UAUNLLY

[even : usual = unevenly : ?]

4	4	4	4	4	4	N	4	4	3	3	2	1	**0**	**0**	**0**
3	**3**	3	3	3	3	E	3	3	3	2	1	**0**	0	0	0
2	**2**	2	2	2	2	V	2	2	2	1	**0**	0	0	0	0
1	**1**	1	1	1	1	E	1	1	1	**0**	0	0	0	0	0
0	**0**	**0**	**0**	**0**	**0**	A◁	**0**	**0**	**0**	0	0	0	0	0	0
L	A	U	S	U	B		C	U	N	E	V	E	N	L	Y

△ △

⟨ε,U,ε, [DEL,•], MOVE B + COPY U ⟩

				E	V	E	N		
U	S	U	A				L		

△

			E	V	E	N			
U	N	E	V	E	N	L	Y		

△

SUAUNLLY

[even : usual = unevenly : ?]

4	4	4	4	4	4	N	4	4	3	3	2	1	**0**	**0**	**0**
3	**3**	3	3	3	3	E	3	3	3	2	1	**0**	0	0	0
2	**2**	2	2	2	2	V	2	2	2	1	**0**	0	0	0	0
1	**1**	1	1	1	1	E	1	1	1	**0**	0	0	0	0	0
0	**0**	**0**	**0**	**0**	**0**	A◁	**0**	**0**	**0**	0	0	0	0	0	0
L	A	U	S	U	B		C	U	N	E	V	E	N	L	Y

△ △

$\langle \varepsilon, \varepsilon, \varepsilon, [\bullet,\bullet], \text{STOP} \rangle$

				E	V	E	N		
U	S	U	A				L		

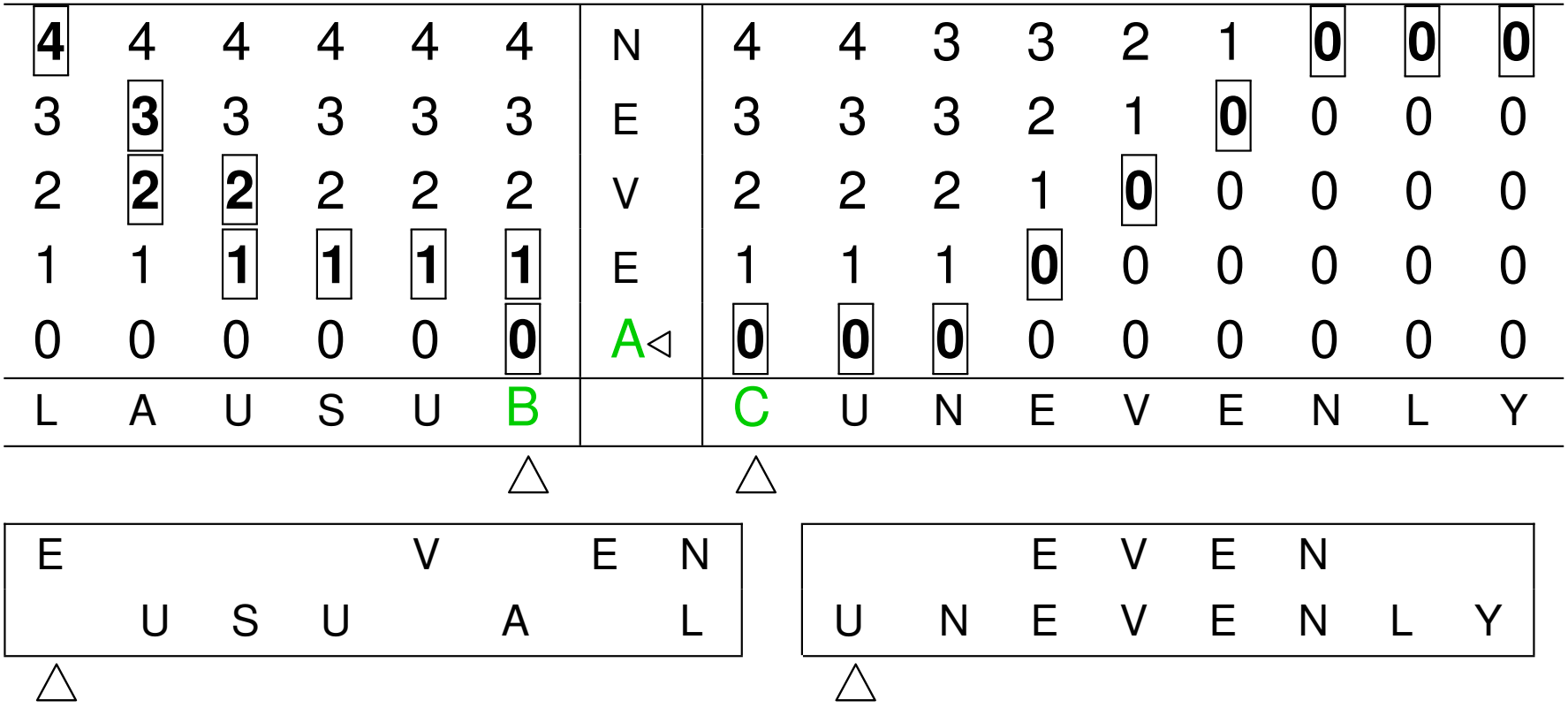
△

		E	V	E	N				
U	N	E	V	E	N	L	Y		

△

USUAUNLLY

[even : usual = unevenly : ?]



UNUSUALLY

[even : usual = unevenly : ?]

15 solutions (681 synchronisations):

uunsually	*usuaunlly*	*usuunalyl*
unusually	*usunually*	*uunslyual*
unulysual	*uunsualyl*	*unuslyual*
unusualyl	*usunualyl*	*uunsulyal*
unusulyal	*usunulyal*	*usuunally*

(72 solutions selon la définition de [Stroppa & Yvon, 2005])

$[\mathtt{x} : \mathtt{y} :: \mathtt{z} : \mathtt{t}] \iff t \in (y \bullet z) \setminus x$

(Yvon & al., 2004)

- **shuffle** $a \bullet b$ lire séquences dans a et b de gauche à droite, en autorisant de changer de chaîne

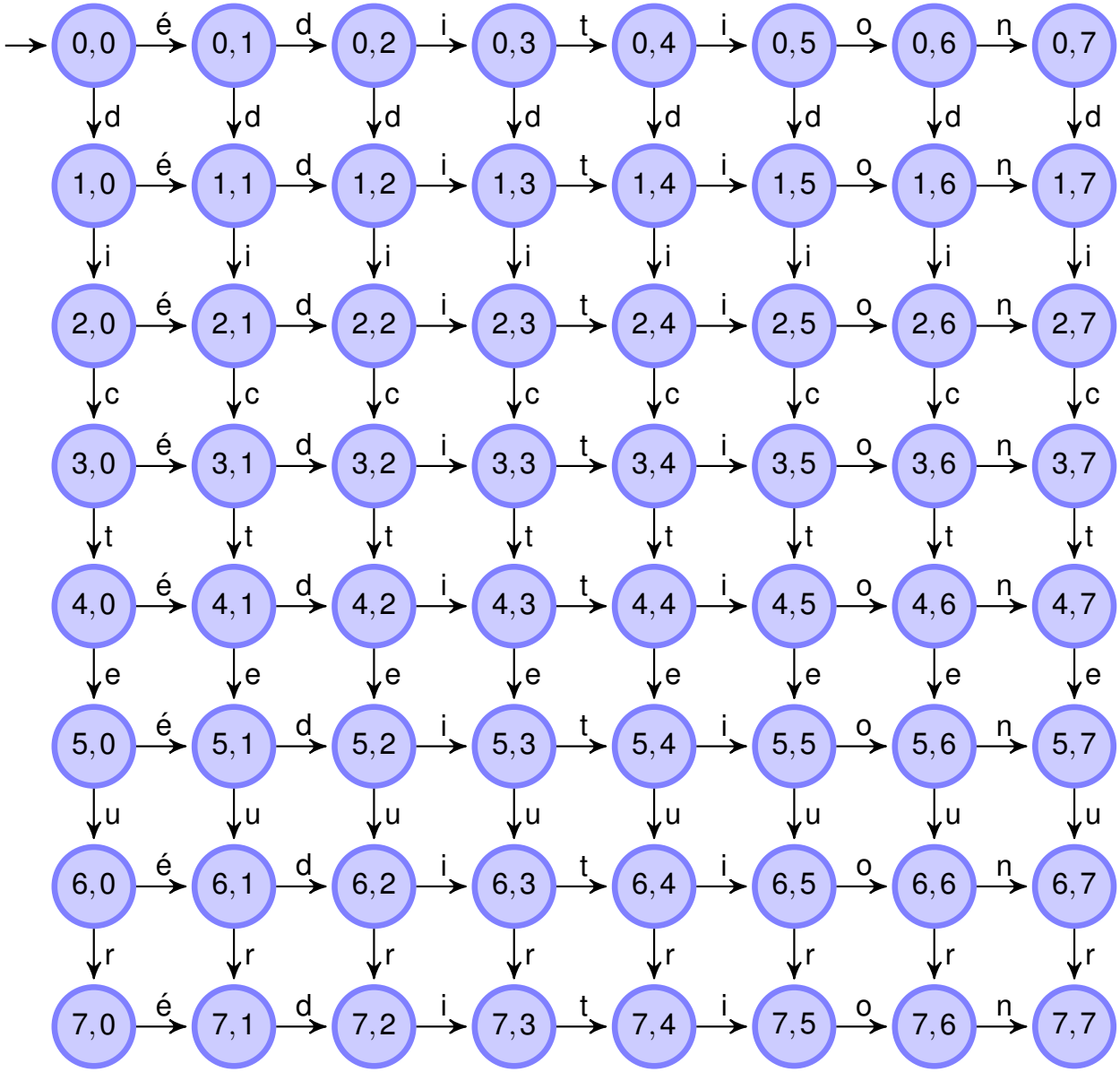
  *spondy*ondonti*lalgia*tis et
  ond*spond*onylal*titis*gia $\in$ *spondylalgia* $\circ$ *ondontitis*

- **complément** $a \setminus b$ chaines obtenues en retirant la sous-chaine b dans a

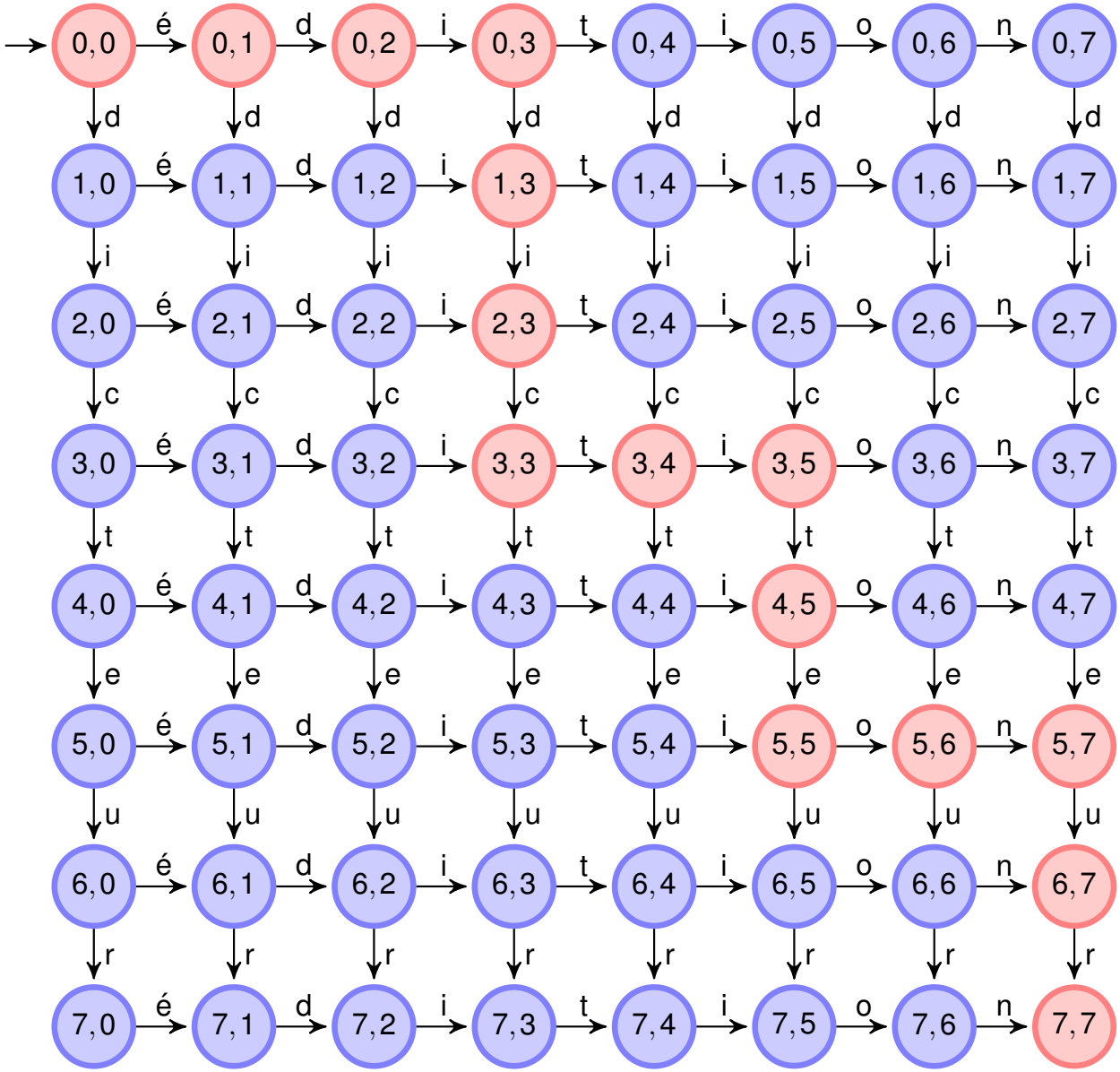
  spondylitis $\in$ *spondy**ondont**il**algia*tis $\setminus$ *ondontalgia*
  spydoniltis $\in$ *sp**ondy**on**dont**il**algia*tis $\setminus$ *ondontalgia*

$\{(y \bullet z) \setminus x\}$ est un langage **régulier**

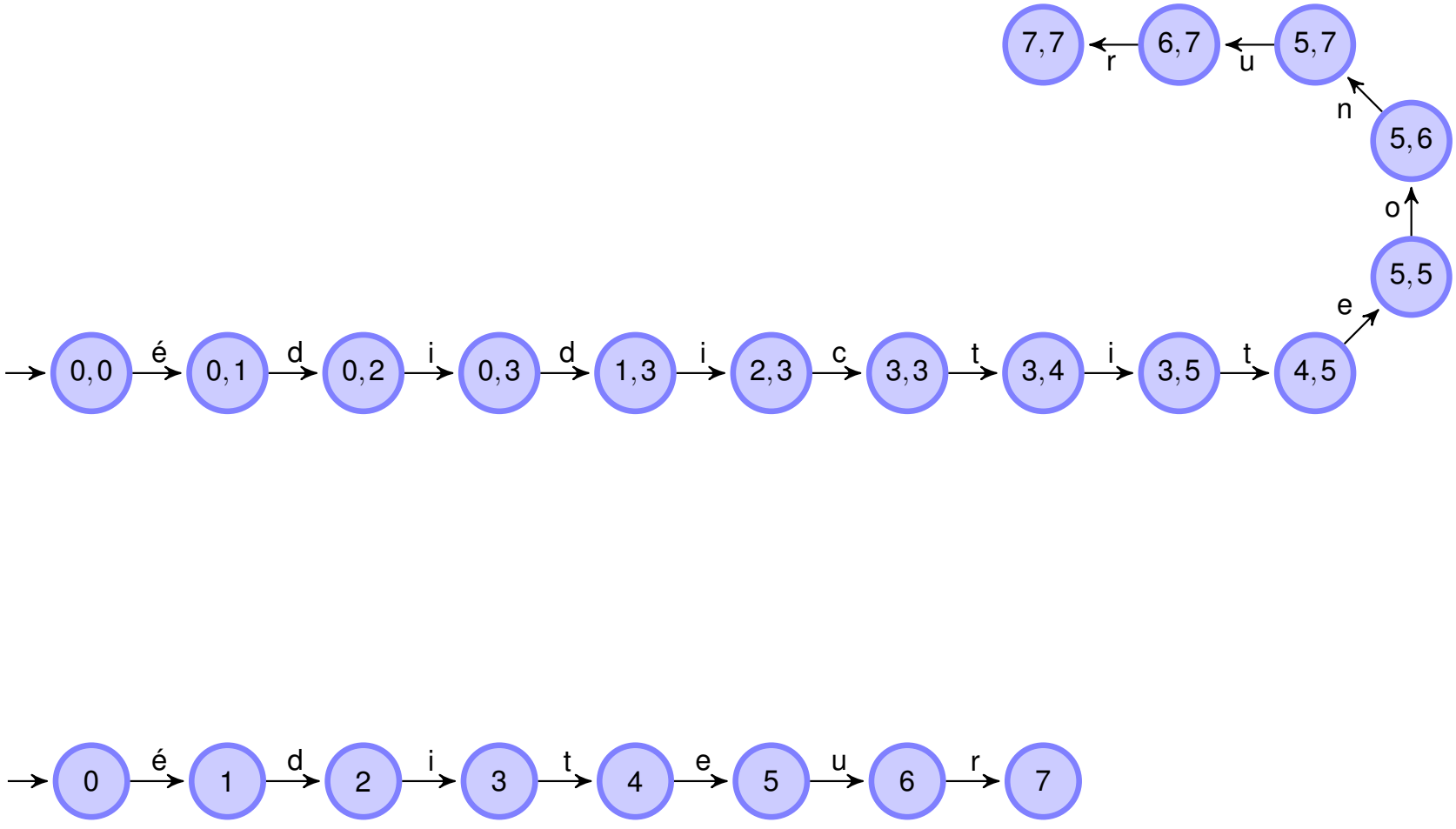
[éditeur : édition :: dicteur : ?]



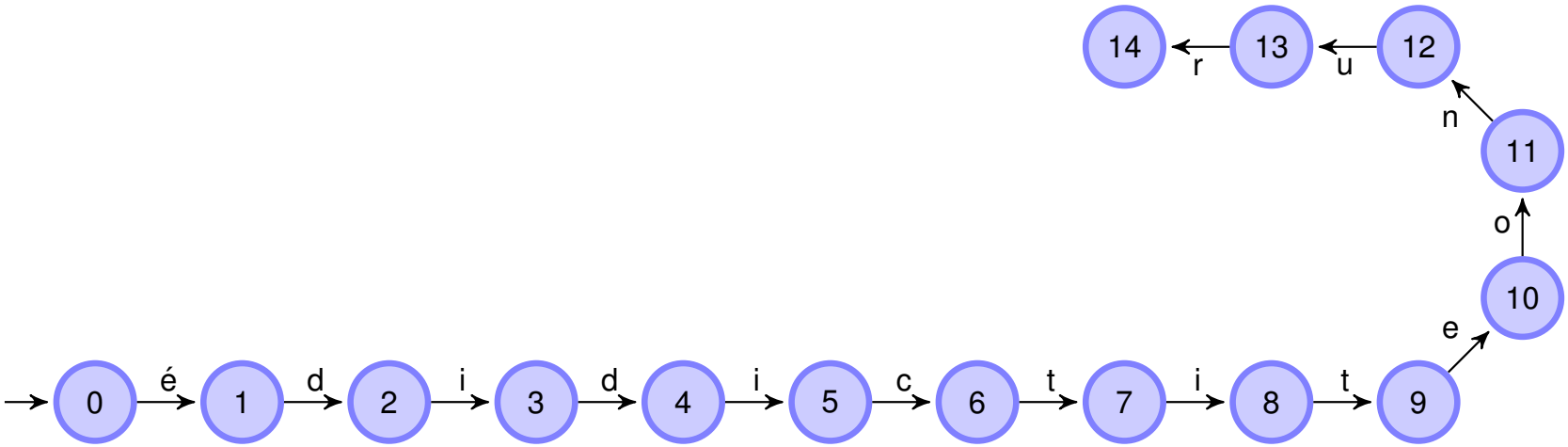
[éditeur : édition :: dicteur : ?]

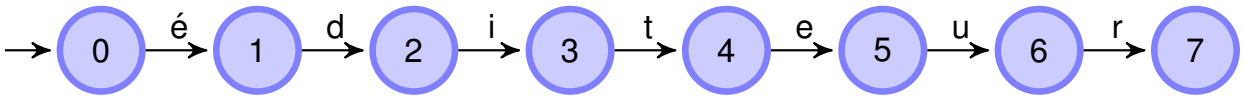


édition • dicteur \ éditeur

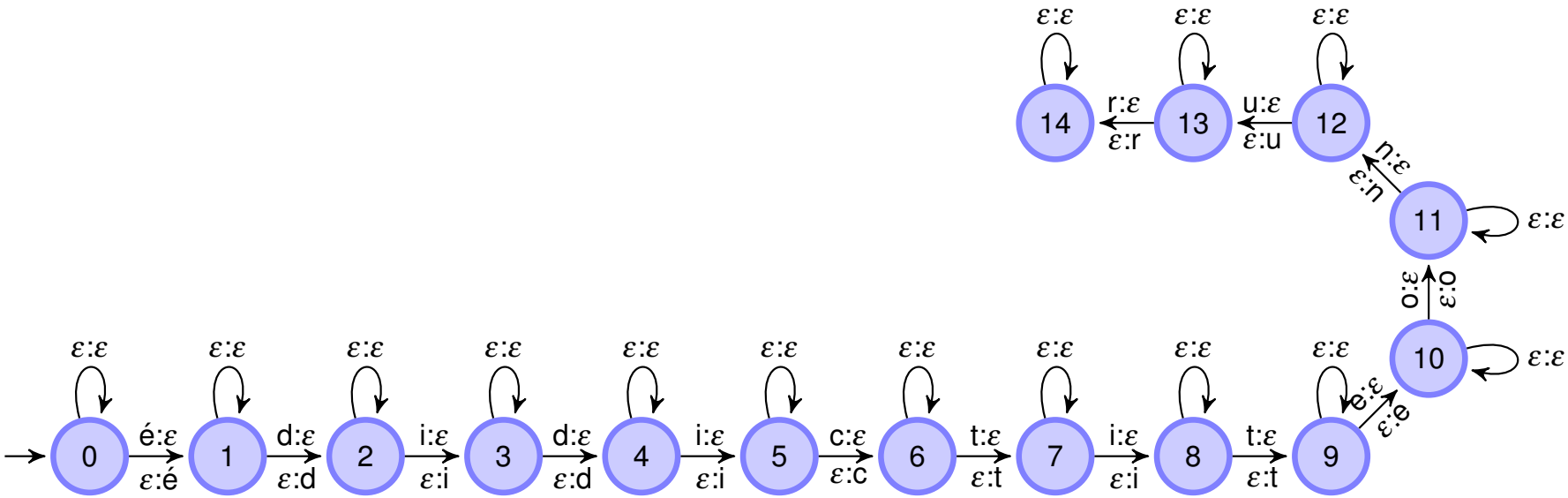


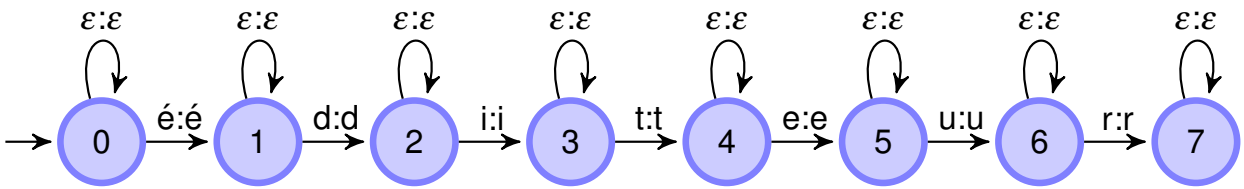
édition • dicteur \ éditeur



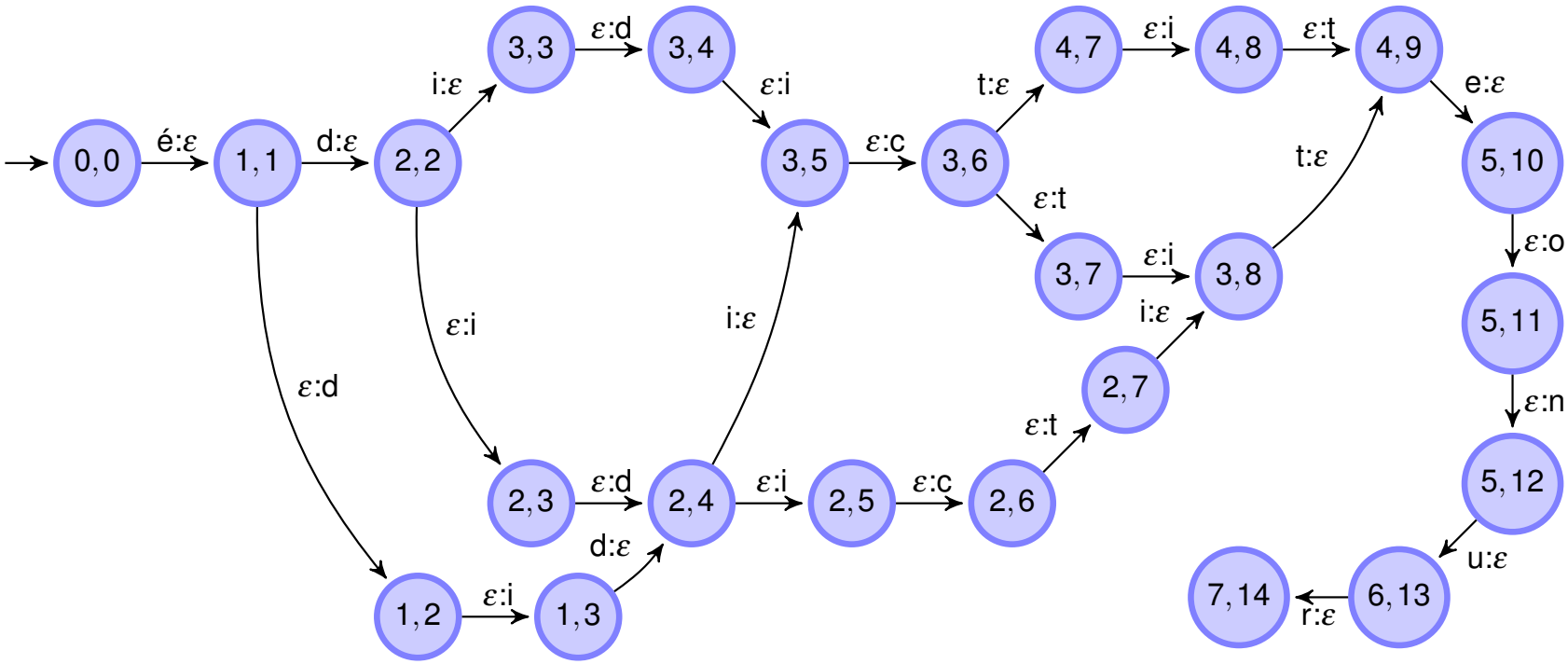


édition • dicteur \ éditeur

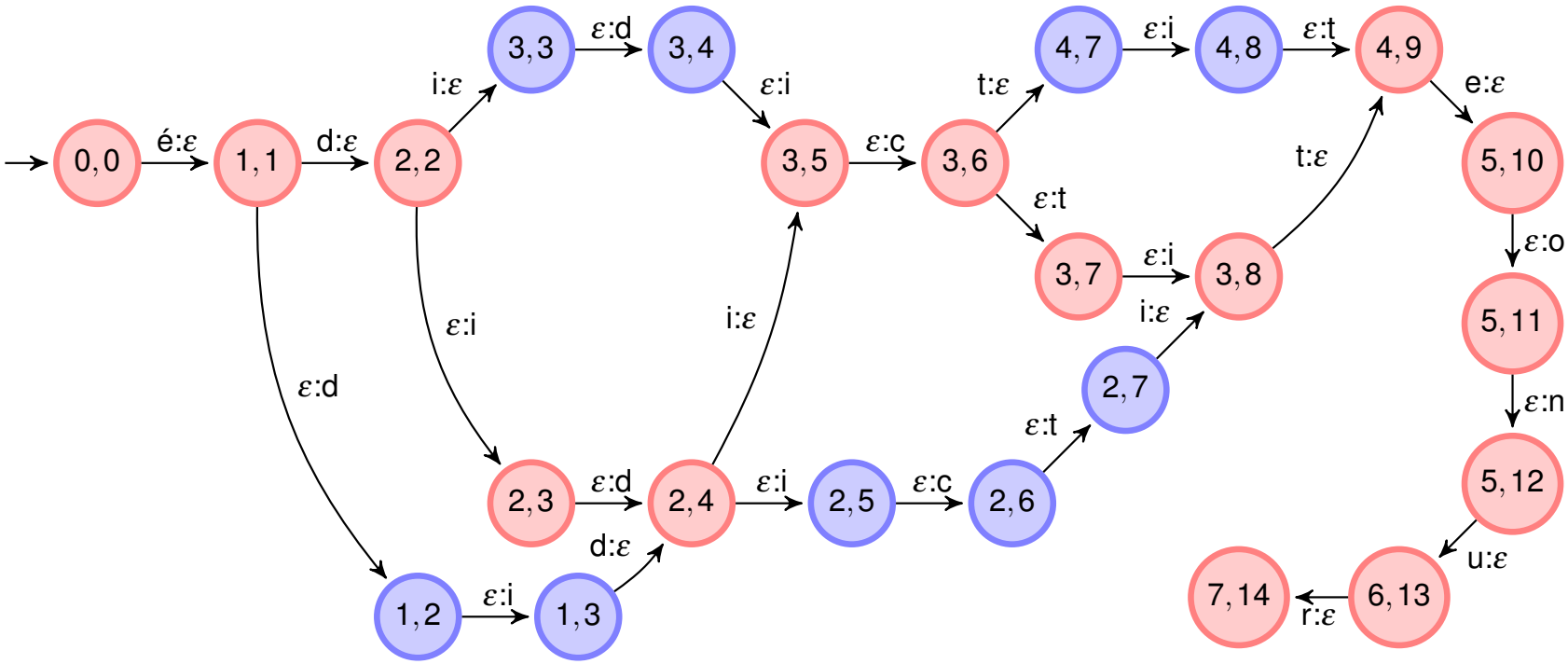




édition • dicteur \ éditeur

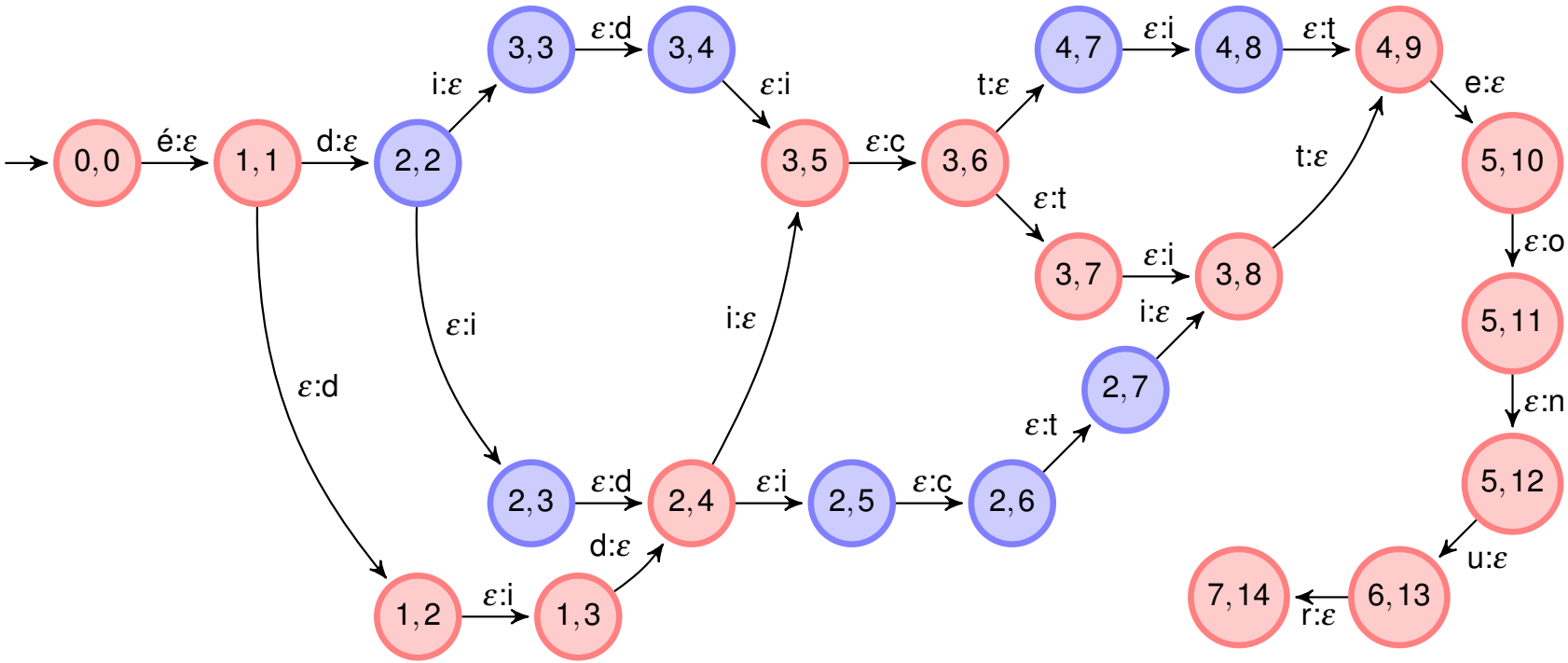


édition • dicteur \ éditeur



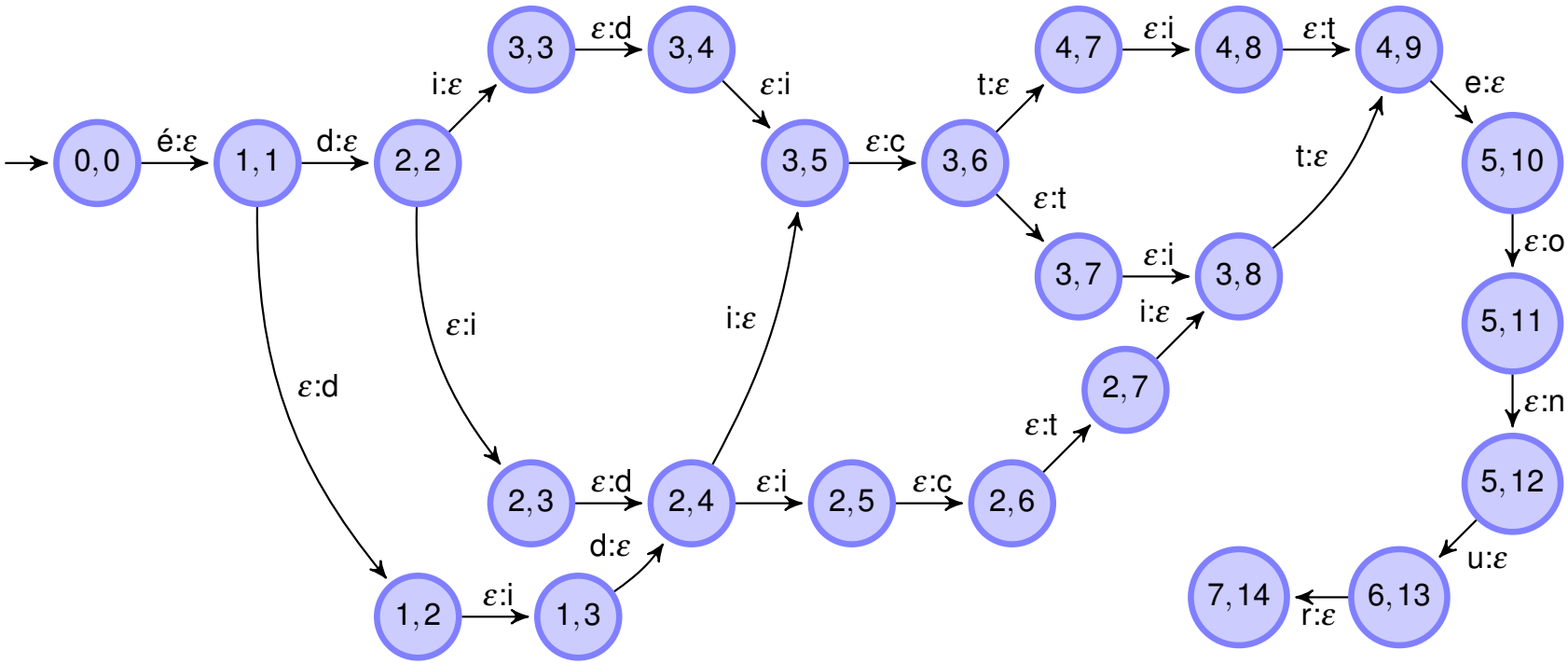
idction

édition • dicteur \ éditeur



idction , diction

édition • dicteur \ éditeur



idction , diction , idciton, diicton, diciton

[éditeur : édition :: dicteur : ?]

110 sol. $f(diction) = 5180, f(ditcion) = 1155, \ldots, f(tiondic) = 1$

diction ditcion dcition diticon diciton diicton diitcon idction
ditiocn diciotn cdition dticion dtiicon diicotn diioctn icdtion
dctiion idtcion idticon itdicon ditionc diciont diitocn idciton
dciiton dtiiocn idicton iditcon diicont diiocnt diionct ditoicn
idtiocn itdcion tdicion dtioicn itdiocn tdiicon dioictn icditon
dciiotn cdtiion ctdiion ictdion idciotn itidcon tidicon diitonc
dtiionc idicotn idioctn cdiiton iditocn itdoicn itidocn tdiiocn
idtionc dioicnt dioinct dionict ditoinc ditonic itdionc dciiont
dtioinc dtionic idciont idtoicn itiodcn itodicn tdioicn tidiocn
tidoicn tiodicn cdiiotn icdiotn idoictn iodictn idicont idiocnt
idionct itidonc iditonc itdoinc itdonic itiodnc tdiionc cdiiont
icdiont idoicnt idoinct idonict idtoinc idtonic iodicnt iodinct
iodnict iondict itiondc itodinc itodnic itondic tdioinc tdionic
tidionc tidoinc tidonic tiodinc tiodnic tiondic

À propos de ces solveurs

- 0, 1 ou plusieurs solutions (souvent inutiles) à une équation analogique

À propos de ces solveurs

- 0, 1 ou plusieurs solutions (souvent inutiles) à une équation analogique
- solutions(lepage) $\subseteq$ solutions(Stroppa&Yvon)

À propos de ces solveurs

- 0, 1 ou plusieurs solutions (souvent inutiles) à une équation analogique
- solutions(lepage) $\subseteq$ solutions(Stroppa&Yvon)
- les deux solveurs ont une complexité exponentielle
  - implémentation par échantillonage
  - aucune garantie d'obtenir une solution lorsqu'il en existe une

Exemple

[even : usual :: unevenly : ?]

ρ	*nb*	solutions		
20	12	usuaunlly (3)	unusually (2)	usunually (2)
100	34	unusually (6)	usuaunlly (6)	uunsually (4)
1000	67	unusually (57)	uunsually (23)	usuunally (19)
2000	72	unusually (130)	uunsually (77)	usunually (43)

- ρ : fréquence d'échantillonnage
- *nb*: nombre de solutions produites

Exemple

[this guy drinks too much : this boat sinks :: those guys drink too much : ?]

$\rho = 20$	$nb = 8$
$t = 0.0003$	$rang = \phi$
thos_boate_sinks (2)	
tho_boatse_sinks (2)	
thoboatse__sinks (2)	

$\rho = 100$	$nb = 28$
$t = 0.001$	$rang = 13$
thoboatse__sinks (2)	
tho_boatse_sinks (2)	
those_sboat_sink (2)	

$\rho = 1000$	$nb = 28$
$t = 0.009$	$rang = 2$
those_boat_ssink (5)	
those_boats_sink (5)	
thoes_tboa_sinks (5)	

$\rho = 10^6$	$nb = 19796$
$t = 3.82$	$rang = 10$
thoes_boat_sinks (2550)	
thoses_boat_sink (1037)	
those_boat_ssink (999)	

- ρ: fréquence d'échantillonnage
- nb: nombre de solutions produites
- $rang$: rand de la solution attendue (ϕ veut dire "pas trouvé")
- t: temps mis par le solver (sec.)

Exemple

[this guy drinks too much : this boat sinks :: those guys drink too much : ?]

$\rho = 20$	$nb = 8$
$t = 0.0003$	$rang = \phi$
thos_boate_sinks (2)	
tho_boatse_sinks (2)	
thoboatse__sinks (2)	

$\rho = 100$	$nb = 28$
$t = 0.001$	$rang = 13$
thoboatse__sinks (2)	
tho_boatse_sinks (2)	
those_sboat_sink (2)	

$\rho = 1000$	$nb = 28$
$t = 0.009$	$rang = 2$
those_boat_ssink (5)	
those_boats_sink (5)	
thoes_tboa_sinks (5)	

$\rho = 10^6$	$nb = 19796$
$t = 3.82$	$rang = 10$
thoes_boat_sinks (2550)	
thoses_boat_sink (1037)	
those_boat_ssink (999)	

- ρ: fréquence d'échantillonnage
- nb: nombre de solutions produites
- $rang$: rand de la solution attendue (ϕ veut dire "pas trouvé")
- t: temps mis par le solver (sec.)

Exemple

[this guy drinks too much : this boat sinks :: those guys drink too much : ?]

$\rho = 20$	$nb = 8$
$t = 0.0003$	$rang = \phi$
thos_boate_sinks (2)	
tho_boatse_sinks (2)	
thoboatse__sinks (2)	

$\rho = 100$	$nb = 28$
$t = 0.001$	$rang = 13$
thoboatse__sinks (2)	
tho_boatse_sinks (2)	
those_sboat_sink (2)	

$\rho = 1000$	$nb = 28$
$t = 0.009$	$rang = 2$
those_boat_ssink (5)	
those_boats_sink (5)	
thoes_tboa_sinks (5)	

$\rho = 10^6$	$nb = 19796$
$t = 3.82$	$rang = 10$
thoes_boat_sinks (2550)	
thoses_boat_sink (1037)	
those_boat_ssink (999)	

- ρ: fréquence d'échantillonnage
- nb: nombre de solutions produites
- $rang$: rand de la solution attendue (ϕ veut dire "pas trouvé")
- t: temps mis par le solver (sec.)

Exemple

[this guy drinks too much : this boat sinks :: those guys drink too much : ?]

$\rho = 20$	$nb = 8$
$t = 0.0003$	$rang = \phi$
thos_boate_sinks (2)	
tho_boatse_sinks (2)	
thoboatse__sinks (2)	

$\rho = 100$	$nb = 28$
$t = 0.001$	$rang = 13$
thoboatse__sinks (2)	
tho_boatse_sinks (2)	
those_sboat_sink (2)	

$\rho = 1000$	$nb = 28$
$t = 0.009$	$rang = 2$
those_boat_ssink (5)	
those_boats_sink (5)	
thoes_tboa_sinks (5)	

$\rho = 10^6$	$nb = 19796$
$t = 3.82$	$rang = 10$
thoes_boat_sinks (2550)	
thoses_boat_sink (1037)	
those_boat_ssink (999)	

- ρ: fréquence d'échantillonnage
- nb: nombre de solutions produites
- $rang$: rand de la solution attendue (ϕ veut dire "pas trouvé")
- t: temps mis par le solver (sec.)

Recherche des analogies

1. $\mathscr{E}_{\mathscr{I}}(u) \equiv \{(s,v,w) \in \mathscr{L}^3 \,|\, [\mathtt{I(s)} : \mathtt{I(v)} :: \mathtt{I(w)} : \mathtt{I(u)}]\}$

- Cubique en $\mathscr{I}$ ...
- Nombre quadratique de résolutions d'équations [Lepage & Denoual, 2005]:
  1. soit $(x,y) \in \mathscr{I}^2$
  2. résoudre $[\mathtt{y} : \mathtt{x} :: \mathtt{t} : ?]$
  3. garder les solutions z qui appartiennent à $\mathscr{I}$

  $\Rightarrow$ triplets (x,y,z)

  Car:

  $[\mathtt{x} : \mathtt{y} :: \mathtt{z} : \mathtt{t}] \Leftrightarrow [\mathtt{y} : \mathtt{x} :: \mathtt{t} : \mathtt{z}]$
- Trop couteux cependant pour des espaces sources modérés

  $\Rightarrow$ échantillonnage de (x,y)

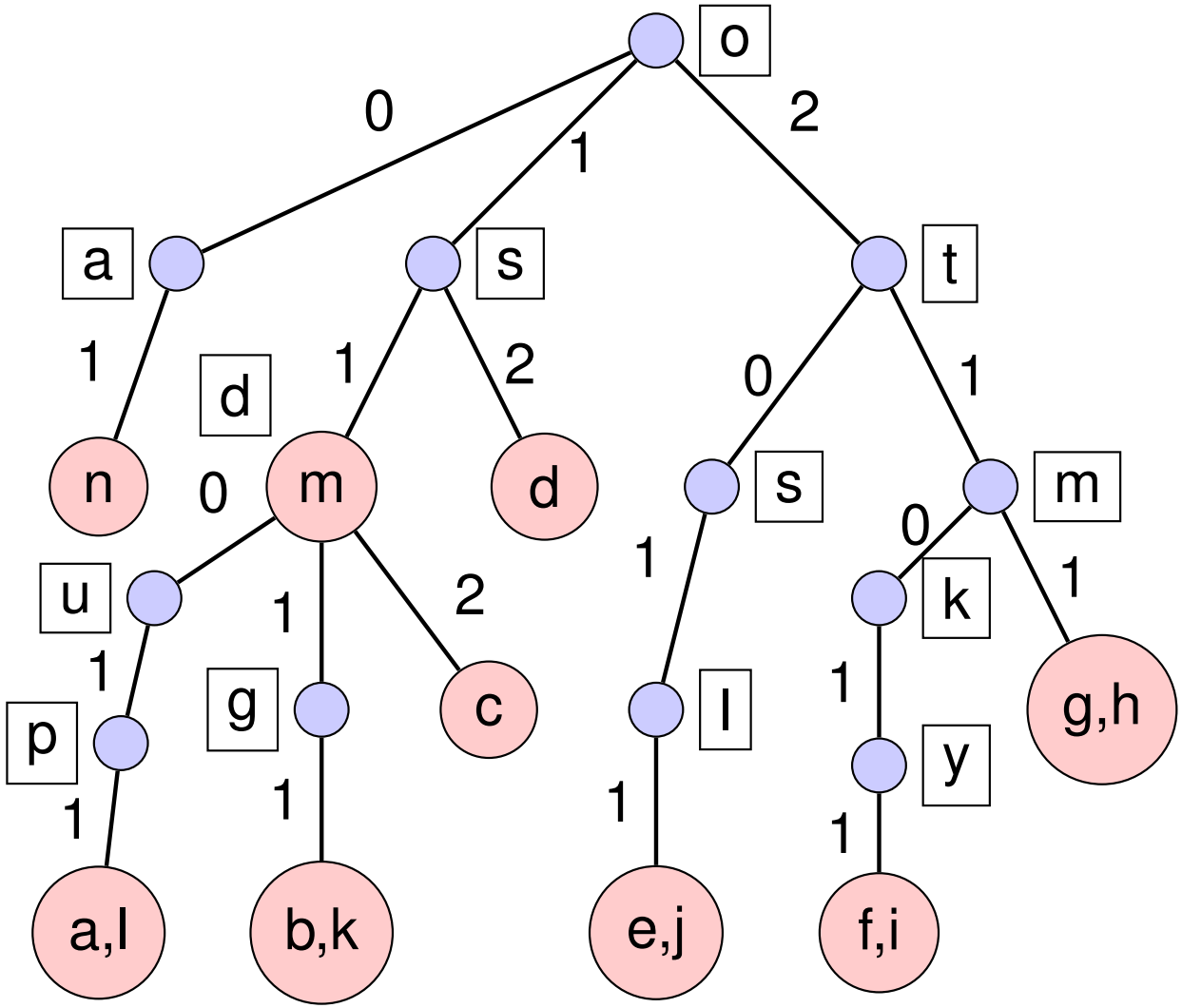
Recherche des analogies

Peut découvrir des **analogies potentielles** en un temps linéaire, grâce à:
$[\mathtt{x} : \mathtt{y} :: \mathtt{z} : \mathtt{t}] \Rightarrow |x|_c + |t|_c = |y|_c + |z|_c \; \forall c \in \mathscr{A}$

1. soit $\mathtt{x} \in \mathscr{I}$
2. chercher les paires (y, z) vérifiant la propriété sur les comptes
3. vérifier les **véritables analogies** (algorithme en $o(|x| \times |y| \times |z| \times |t|)$)

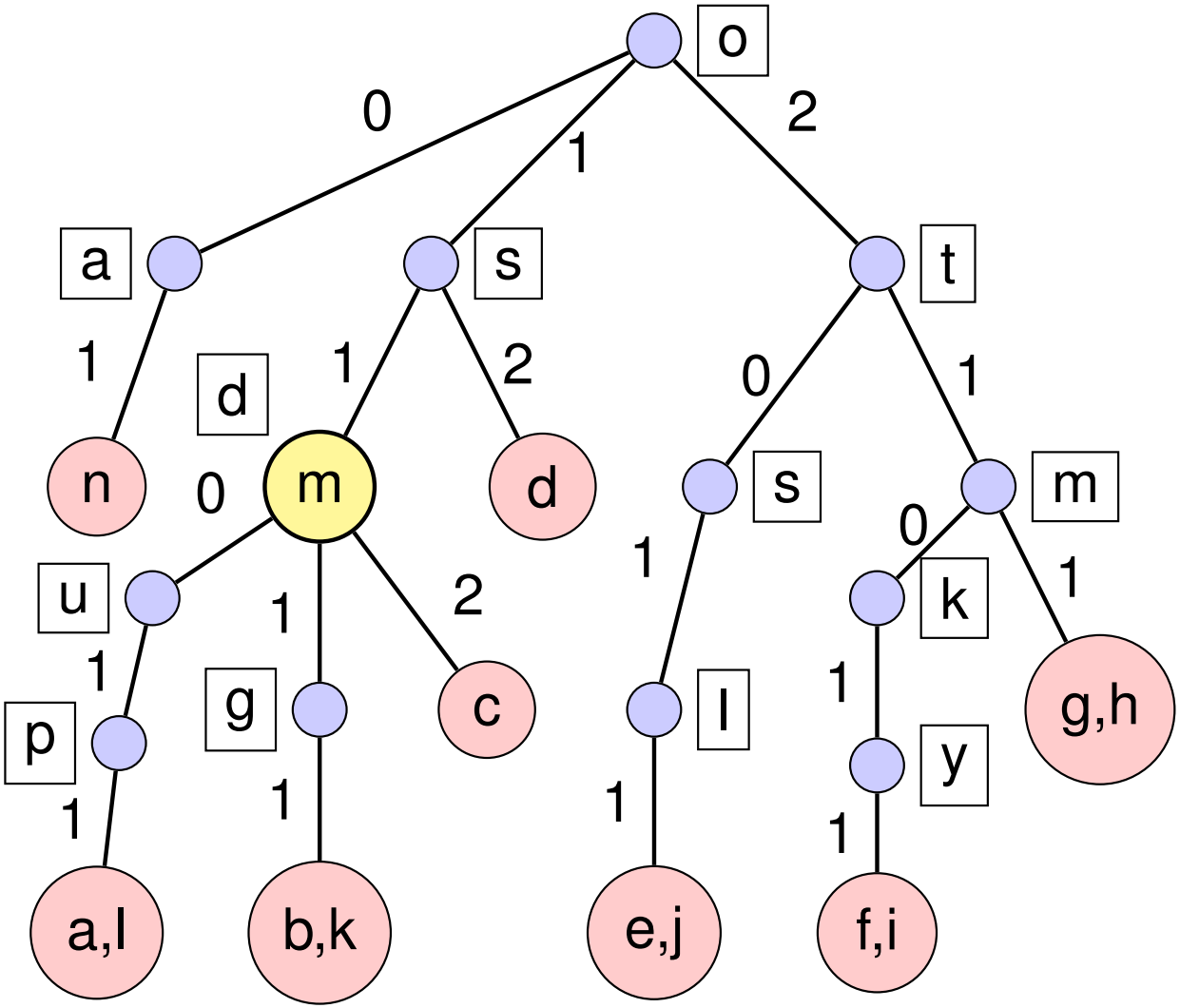
L'espace d'entrée doit être organisé (**tree-count**).

Un tree-count



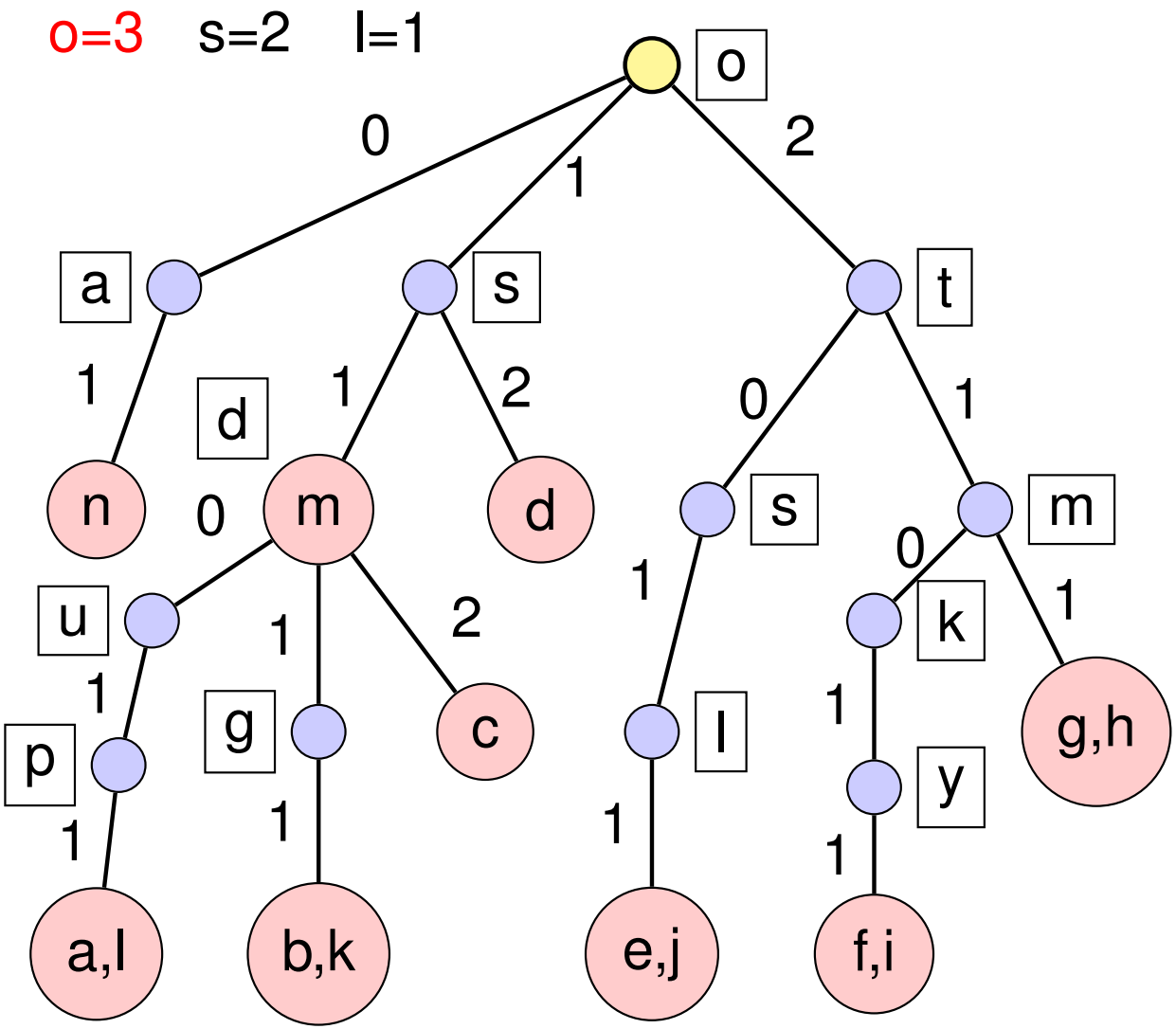
a	soup
b	gods
c	odds
d	sos
e	solo
f	tokyo
g	moot
h	moto
i	kyoto
j	oslo
k	dogs
l	opus
m	os
n	a

Un tree-count



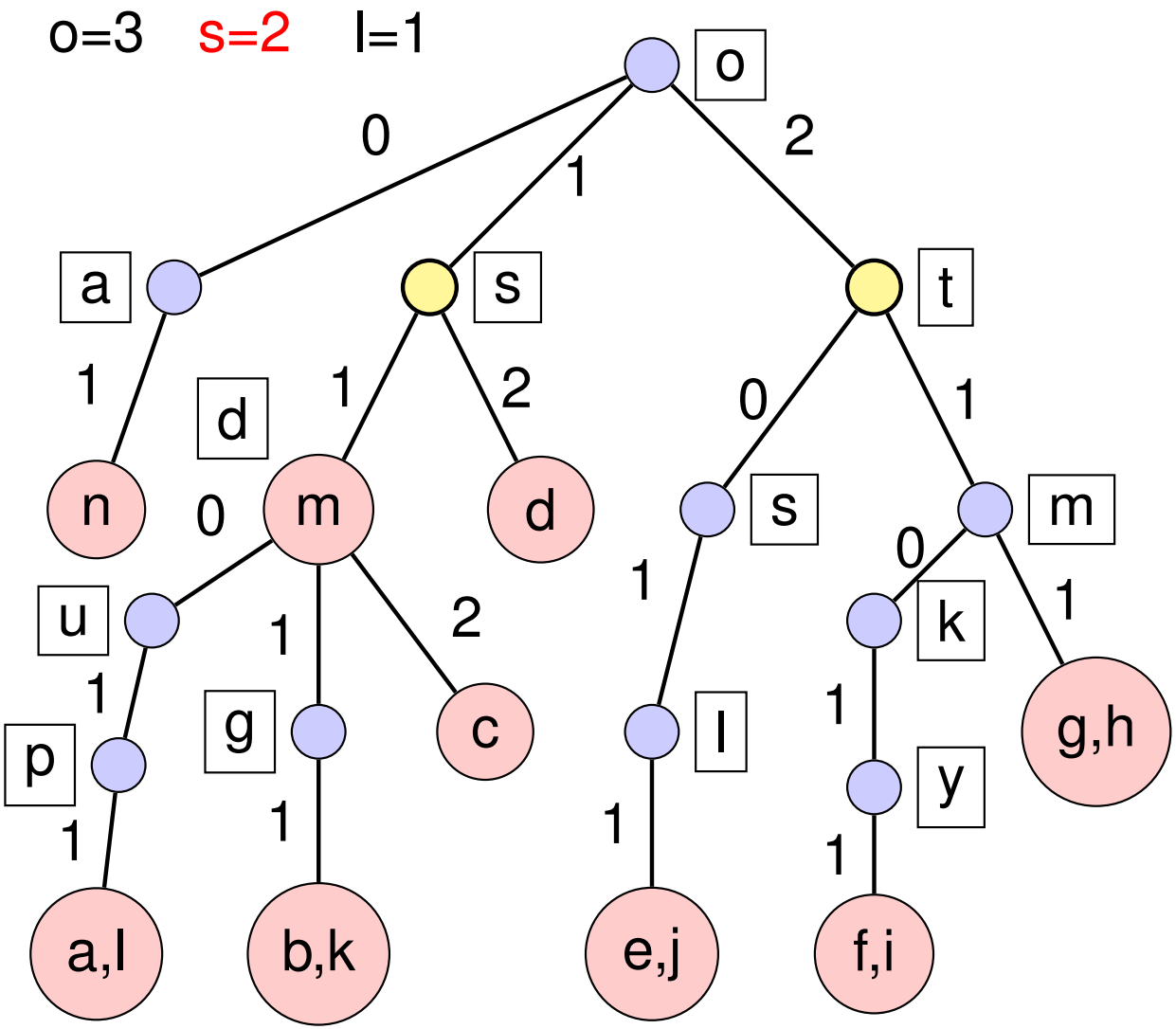
a	soup
b	gods
c	odds
d	sos
e	solo
f	tokyo
g	moot
h	moto
i	kyoto
j	oslo
k	dogs
l	opus
m	os
n	a

Un tree-count



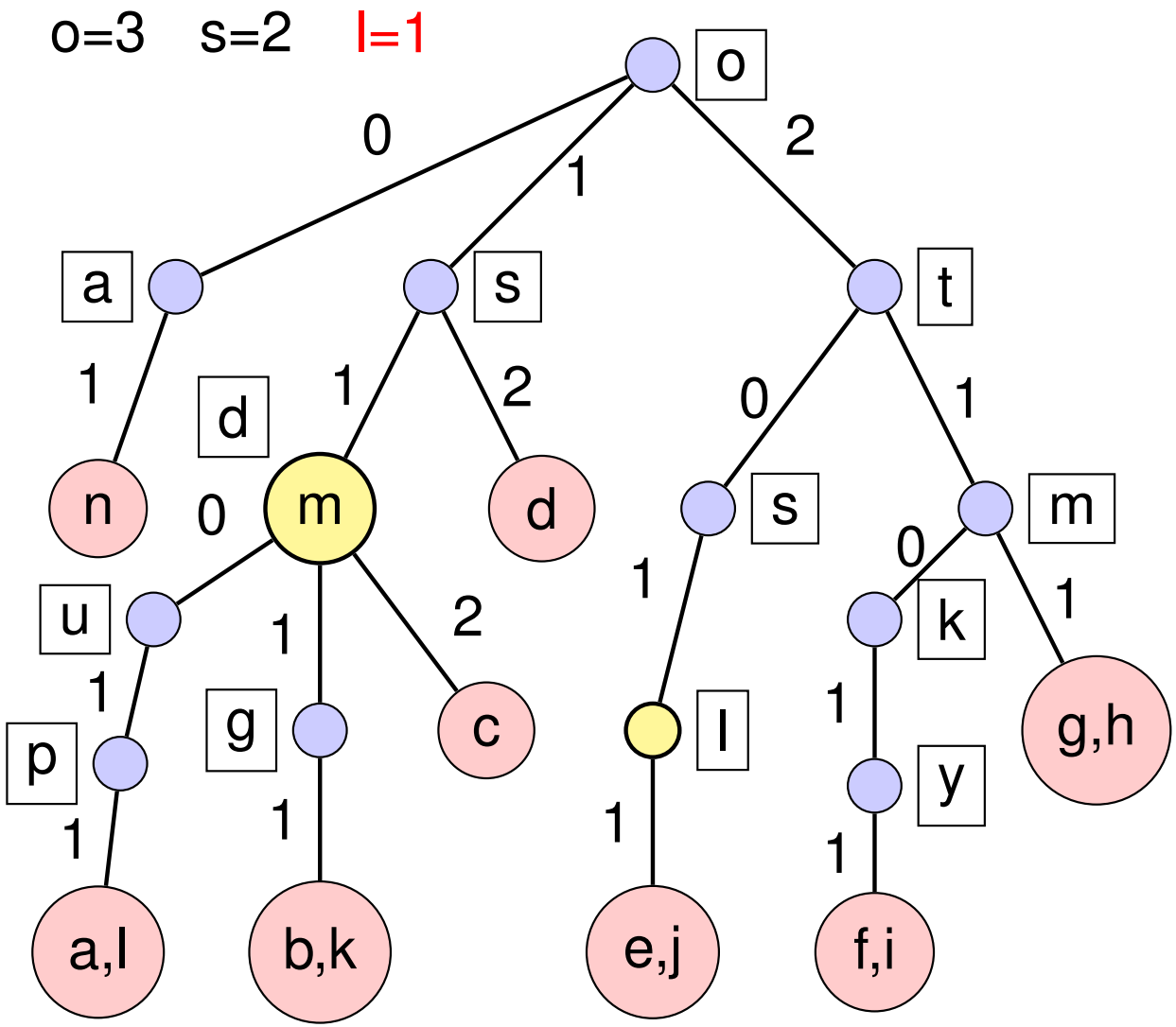
a	soup
b	gods
c	odds
d	sos
e	solo
f	tokyo
g	moot
h	moto
i	kyoto
j	oslo
k	dogs
l	opus
m	os
n	a

Un tree-count



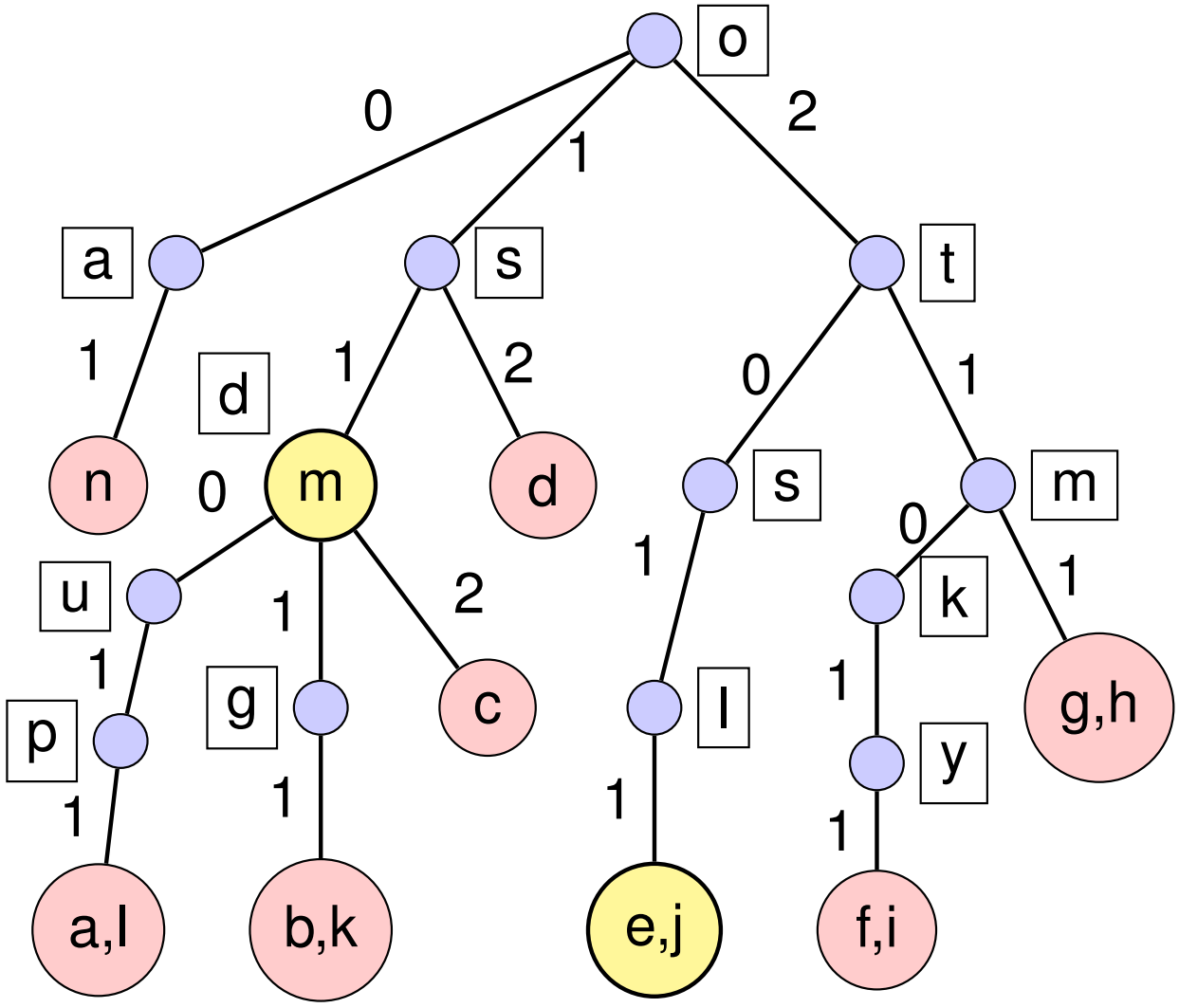
a	soup
b	gods
c	odds
d	sos
e	solo
f	tokyo
g	moot
h	moto
i	kyoto
j	oslo
k	dogs
l	opus
m	os
n	a

Un tree-count



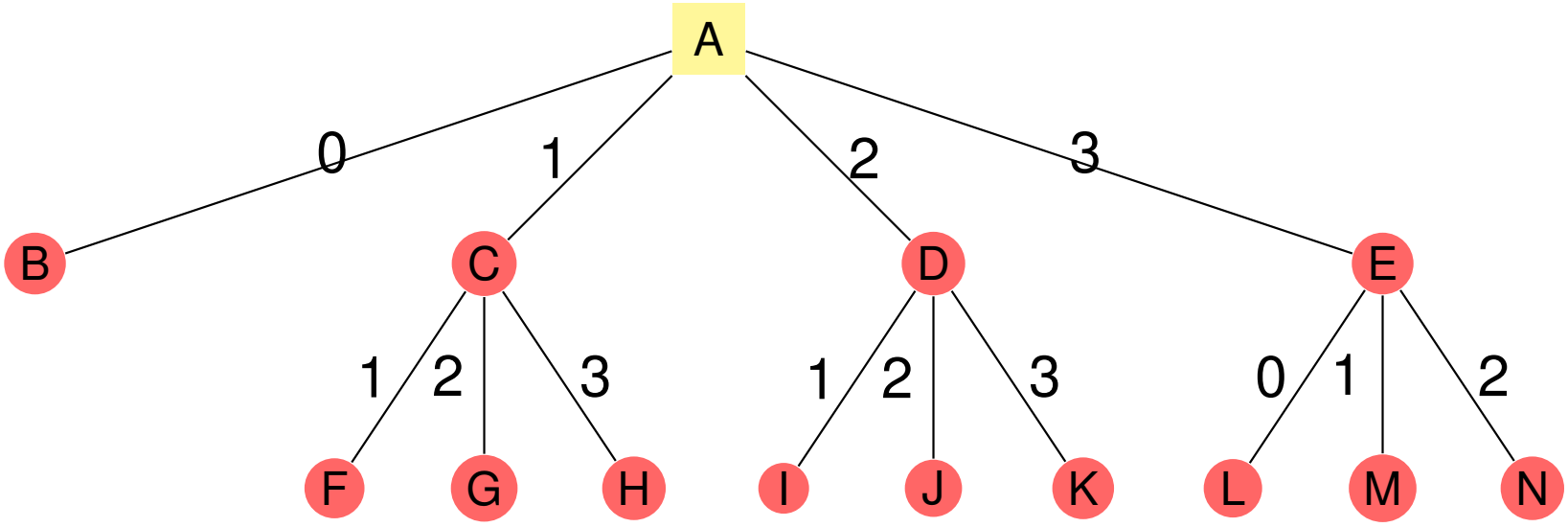
a	soup
b	gods
c	odds
d	sos
e	solo
f	tokyo
g	moot
h	moto
i	kyoto
j	oslo
k	dogs
l	opus
m	os
n	a

Un tree-count



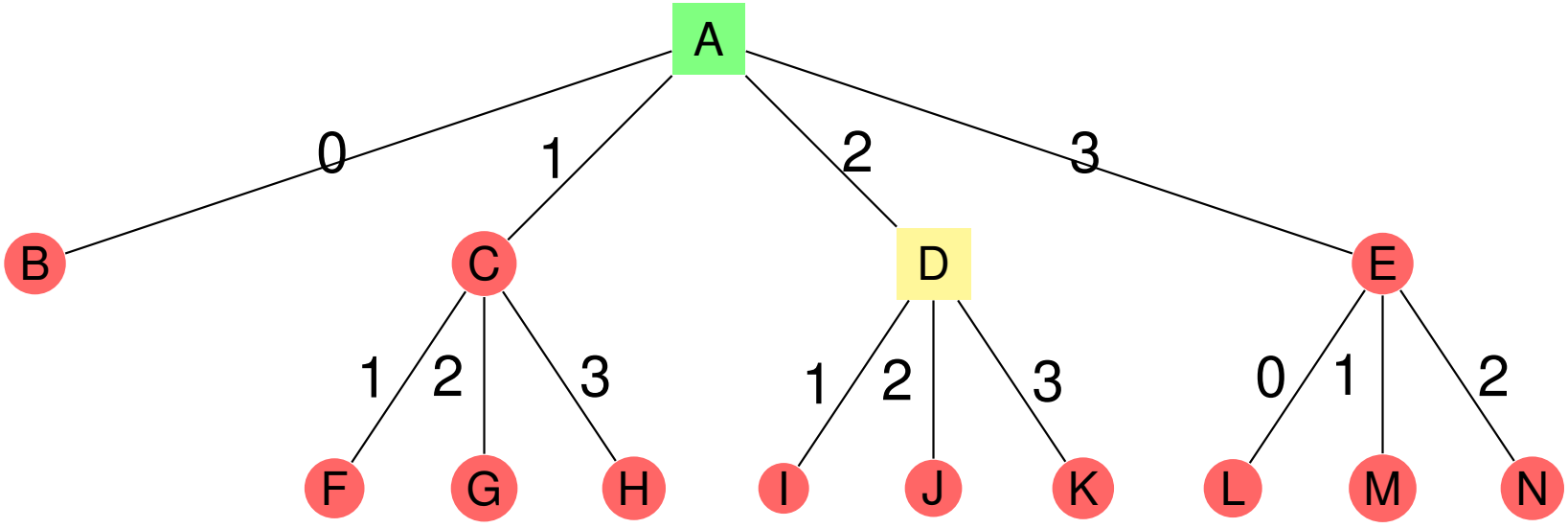
a	soup
b	gods
c	odds
d	sos
e	**solo**
f	tokyo
g	moot
h	moto
i	kyoto
j	**oslo**
k	dogs
l	opus
m	**os**
n	a

Recherche des analogies potentielles



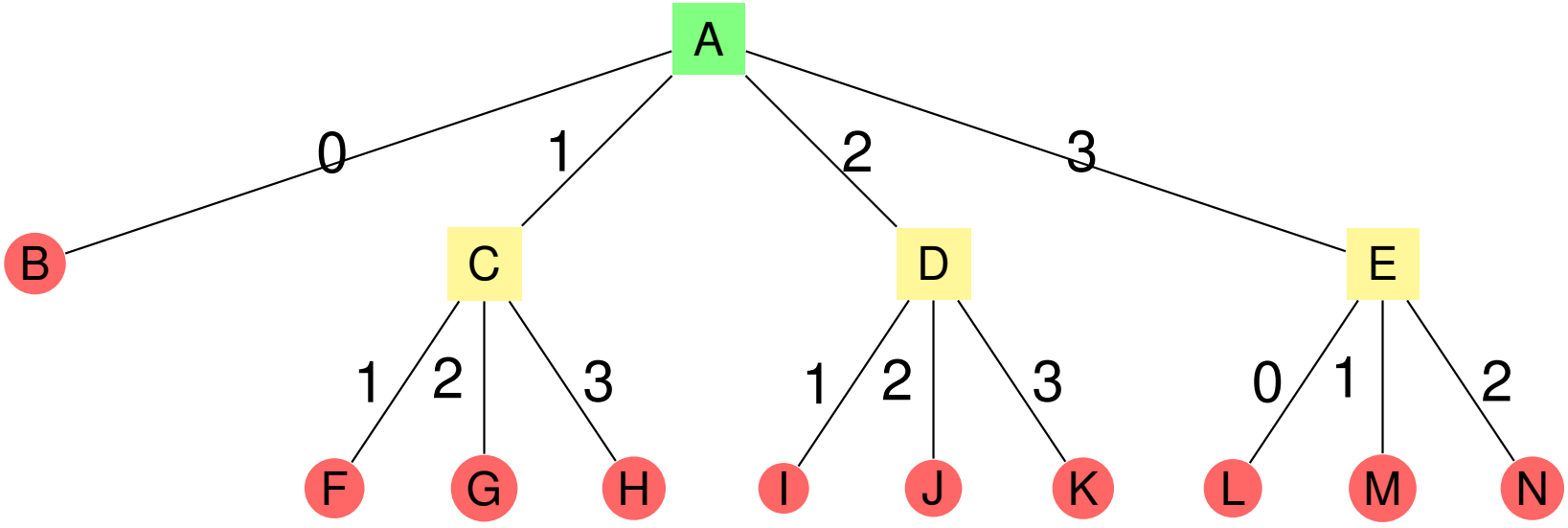
level 0 (A,A)

Recherche des analogies potentielles



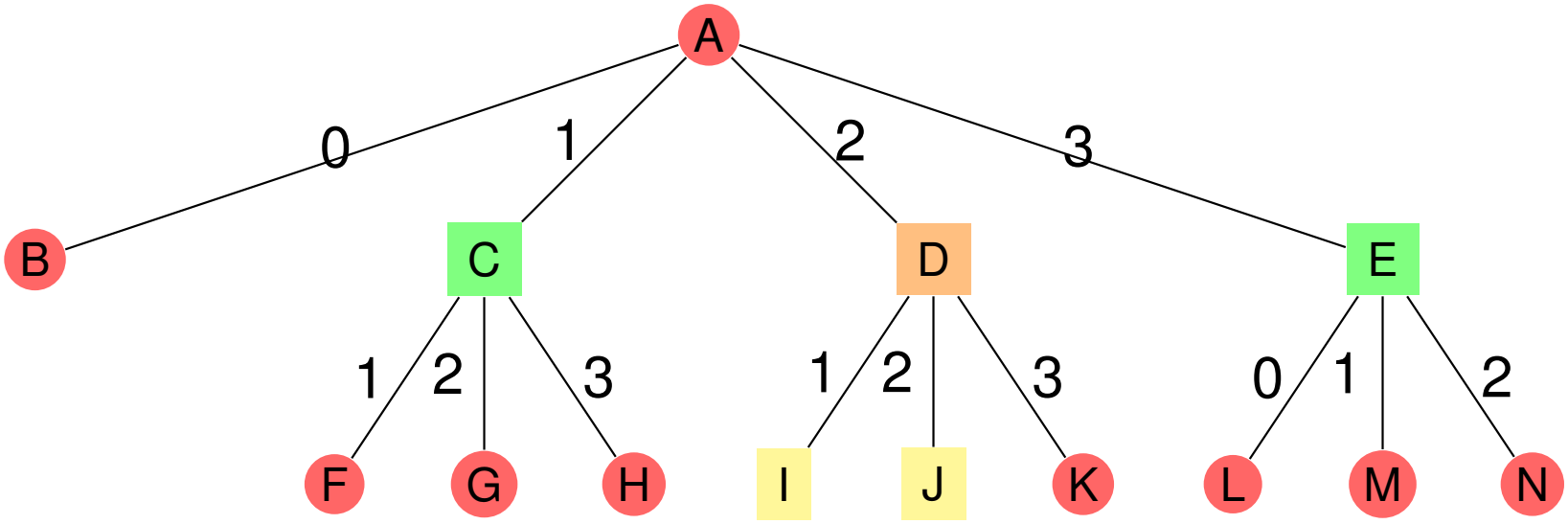
level 0		(A,A)
level 1	4	(D,D)
level 2	3	

Recherche des analogies potentielles



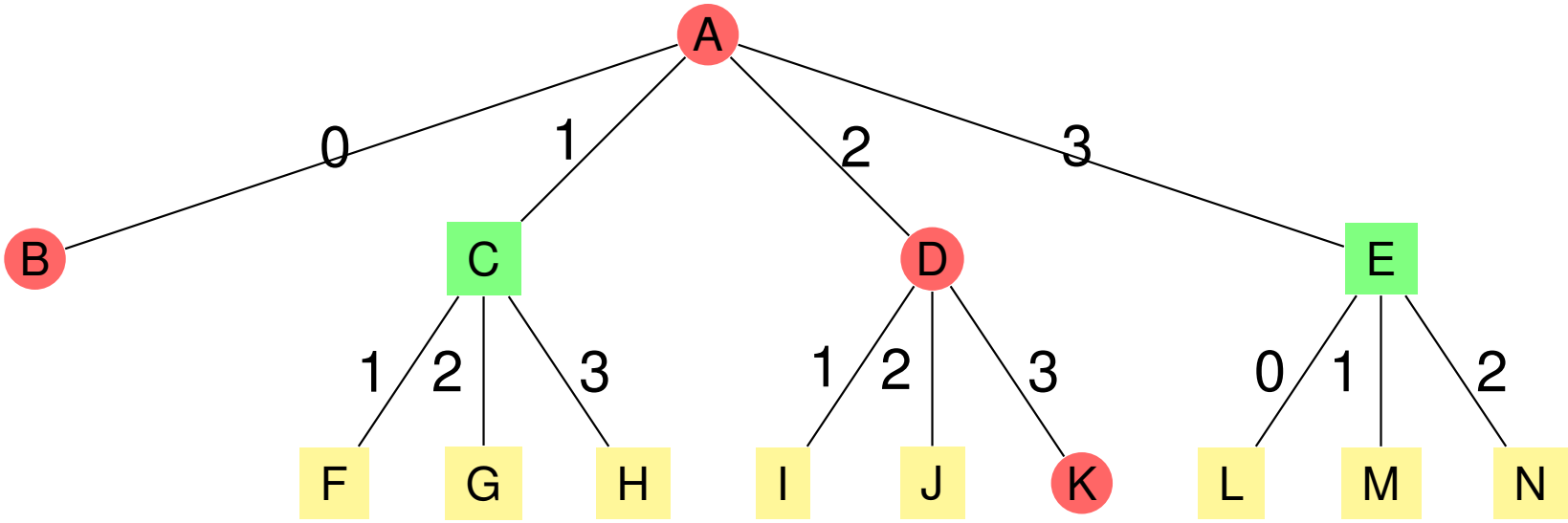
level 0		(A,A)
level 1	4	(D,D) (C,E)
level 2	3	

Recherche des analogies potentielles



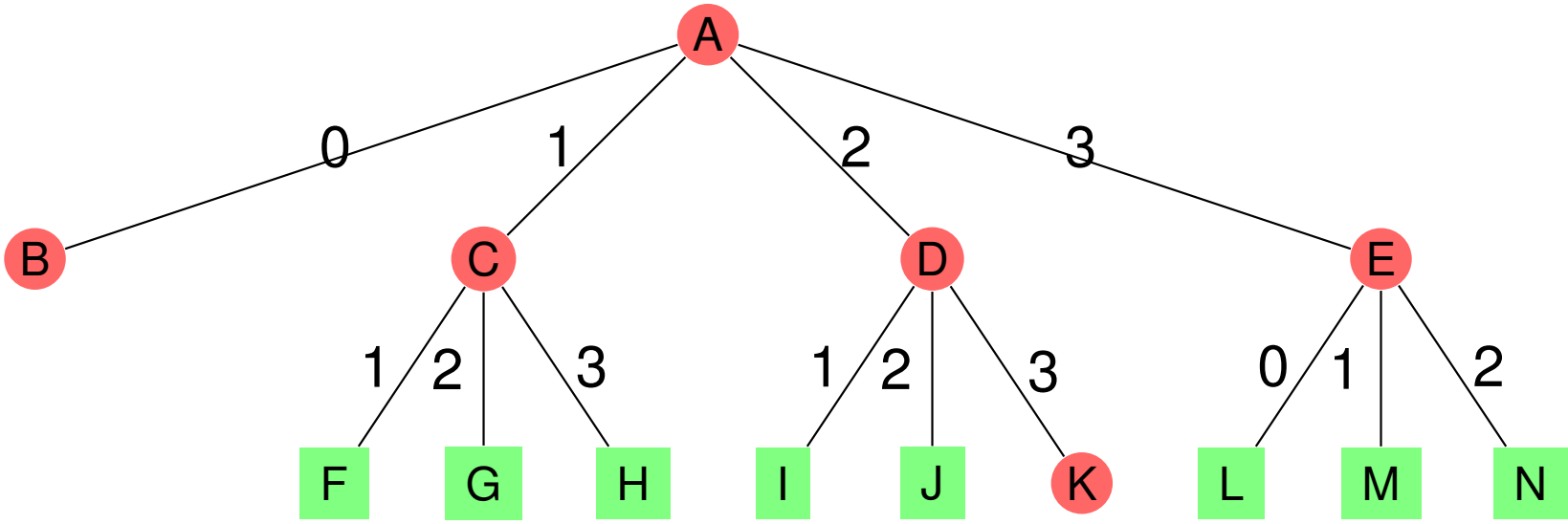
level 0		(A,A)
level 1	4	(D,D) (C,E)
level 2	3	(I,J)

Recherche des analogies potentielles



level 0		(A,A)
level 1	4	(D,D) (C,E)
level 2	3	(I,J) (G,M) (H,L) (F,N)

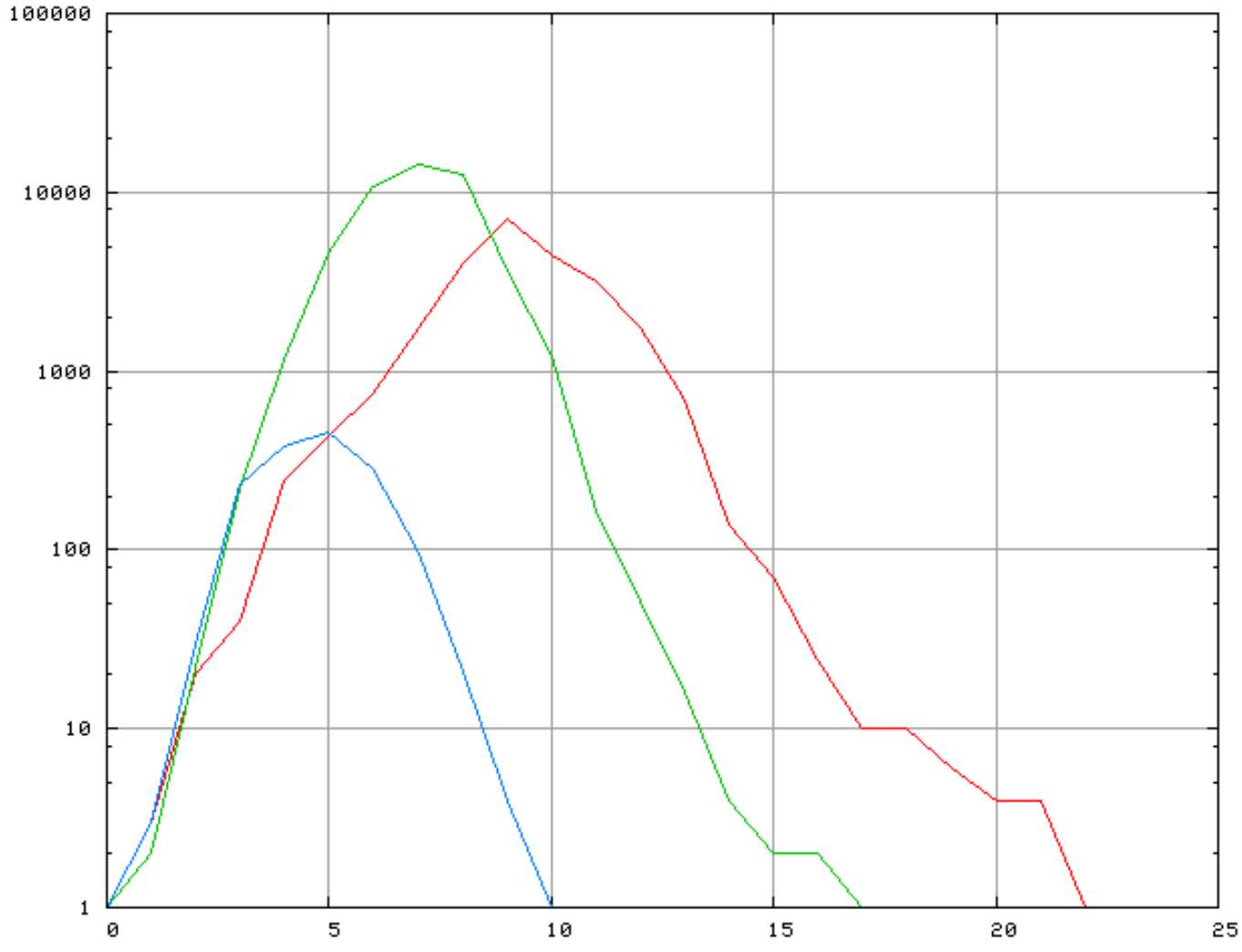
Recherche des analogies potentielles



level 0		(A,A)
level 1	4	(D,D) (C,E)
level 2	3	(I,J) (G,M) (H,L) (F,N)
...		

Recherche dans un tree-count

une figure sans légende . . .



A tree-count

- espace d'entrée: 11 317 717 formes
- en moyenne sur 1 000 recherches:

ratio	time (ms)	$\lvert frontier \rvert$	$\lvert nodes \rvert$
1/1000	5.5e-05	38	6.8
1/100	0.0003	150	6.3
1/10	0.003	1082	6.6
1/5	0.0055	1655	6.5
1/1	0.02	3921	5.8

- mémoire et temps $\sim$ linéaire avec la taille de l'espace d'entrée

Vérification d'une analogie

(Stroppa, 2005)

a(i,j,k,l) $\leftarrow$ **false** , **if** i, j, k or l < 0
for i $\leftarrow$ 0 **to** $|\mathtt{x}|$ **do**
  for j $\leftarrow$ 0 **to** $|\mathtt{y}|$ **do**
    for k $\leftarrow$ 0 **to** $|\mathtt{z}|$ **do**
      for l $\leftarrow$ 0 **to** $|\mathtt{t}|$ **do**
        if i = j = k = l **then**
          a(i,j,k,l) $\leftarrow$ **true**
        else
          a(i,j,k,l) $\leftarrow$

$$\text{or} \begin{cases} \text{a(i-1,j-1,k,l)} \wedge \mathtt{x}[i] = \mathtt{y}[j] \\ \text{a(i-1,j,k-1,l)} \wedge \mathtt{x}[i] = \mathtt{z}[k] \\ \text{a(i,j-1,k,l-1)} \wedge \mathtt{t}[l] = \mathtt{y}[j] \\ \text{a(i,j,k-1,l-1)} \wedge \mathtt{t}[l] = \mathtt{z}[k] \end{cases}$$

return a($|\mathtt{x}|, |\mathtt{y}|, |\mathtt{z}|, |\mathtt{t}|$)

Vérification d'une analogie

- trop long . . .
- vérifier d'abord ($\sim$ 35% d'économie):

$$
\begin{aligned}
&[\mathtt{x} : \mathtt{y} :: \mathtt{z} : \mathtt{t}] \Rightarrow \\
&\quad (\mathtt{x}[1] \in \{\mathtt{y}[1], \mathtt{z}[1]\}) \vee (\mathtt{t}[1] \in \{\mathtt{y}[1], \mathtt{z}[1]\}) \\
&\quad (\mathtt{x}[\$] \in \{\mathtt{y}[\$], \mathtt{z}[\$]\}) \vee (\mathtt{t}[\$] \in \{\mathtt{y}[\$], \mathtt{z}[\$]\})
\end{aligned}
$$

Espaces d'entrée très grands

- 0.02 ms $\times$ 10M = 200 seconds
  (sans même compter la vérification d'analogie)
  $\Rightarrow$ échantillonnage des $\mathtt{x}$-forms ...

- Step 1:
  1. **sample** $\mathtt{x}$-forms $\in \mathscr{I}$
  2. search for all the pairs (y, z) vérifiant la propriété des comptes
  3. vérifier les **véritables analogies**
     (algorithme en $o(|x| \times |y| \times |z| \times |t|)$)

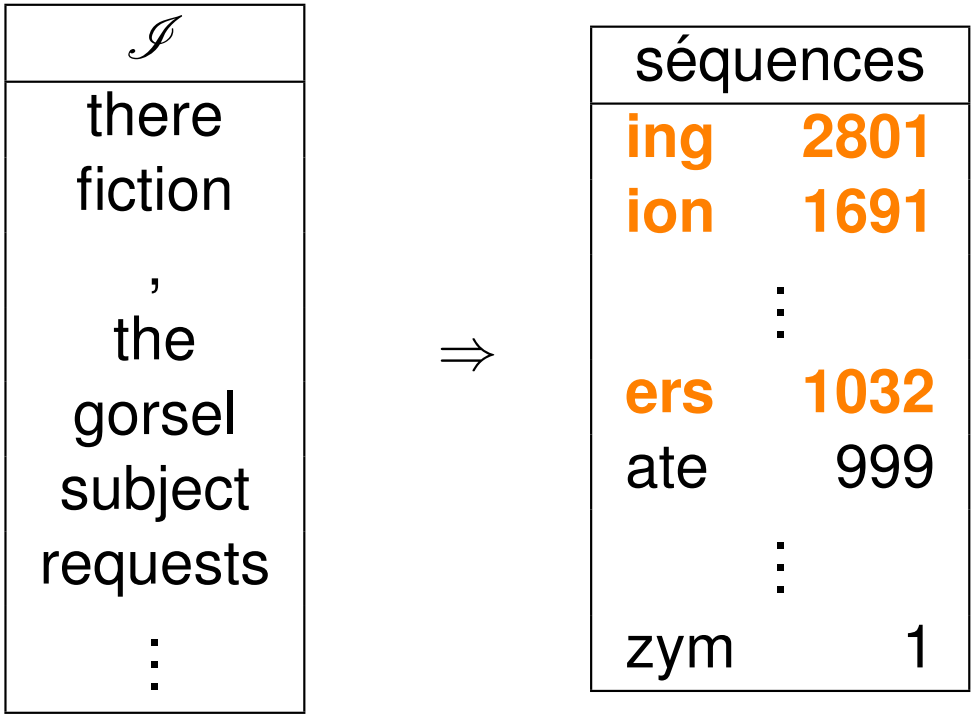
Espaces d'entrée très grands

- Échantillonnage des `x`-forms
  - [une pomme verte : des pommes vertes :: une voiture rouge : des voitures rouges]
  - [dream : dreamer :: dreams : dreamers]
  - [This guy drinks too much : This boat sinks :: These guys drank too much : These boats sank]
- l'approche vectorielle semble adaptée

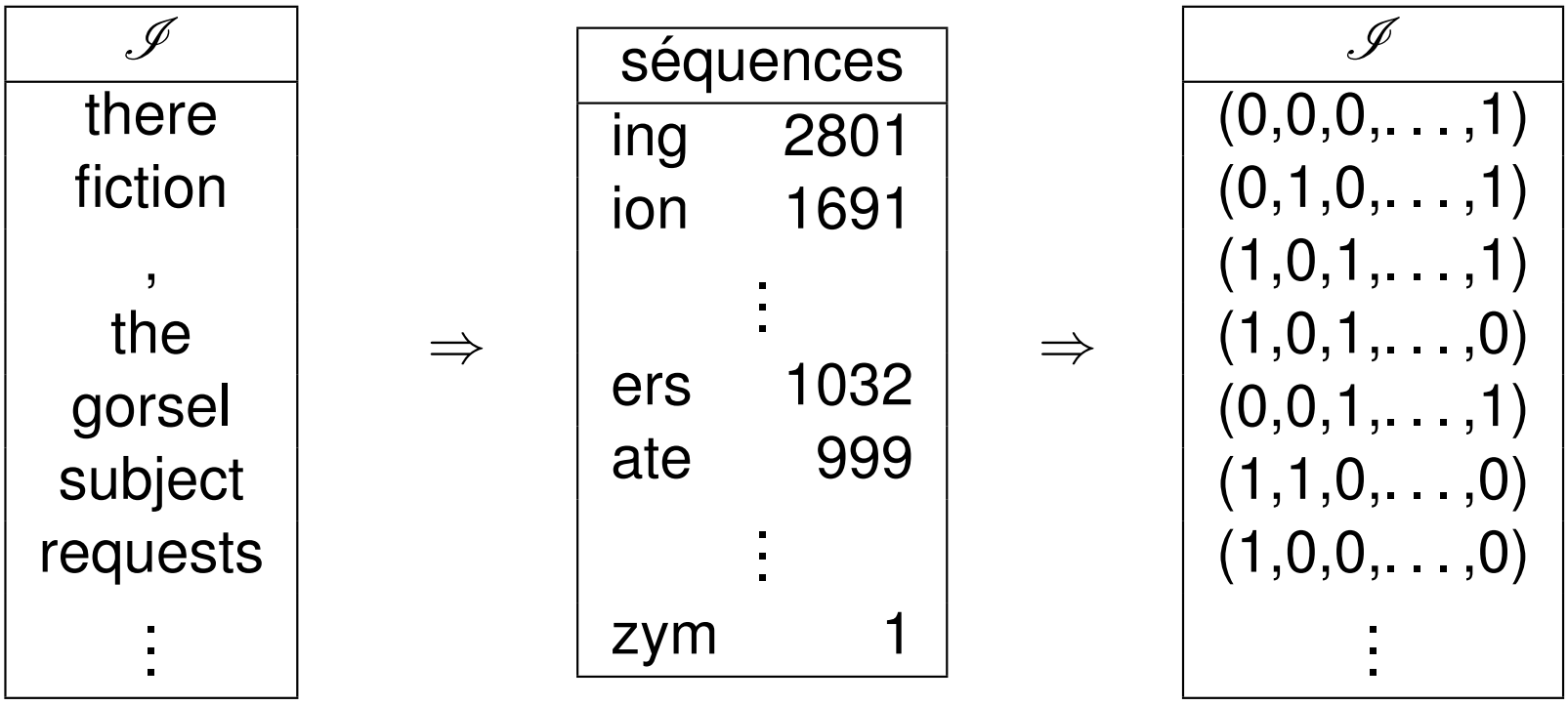
Vector space sampling

$\mathscr{I}$
there
fiction
,
the
gorsel
subject
requests
⋮

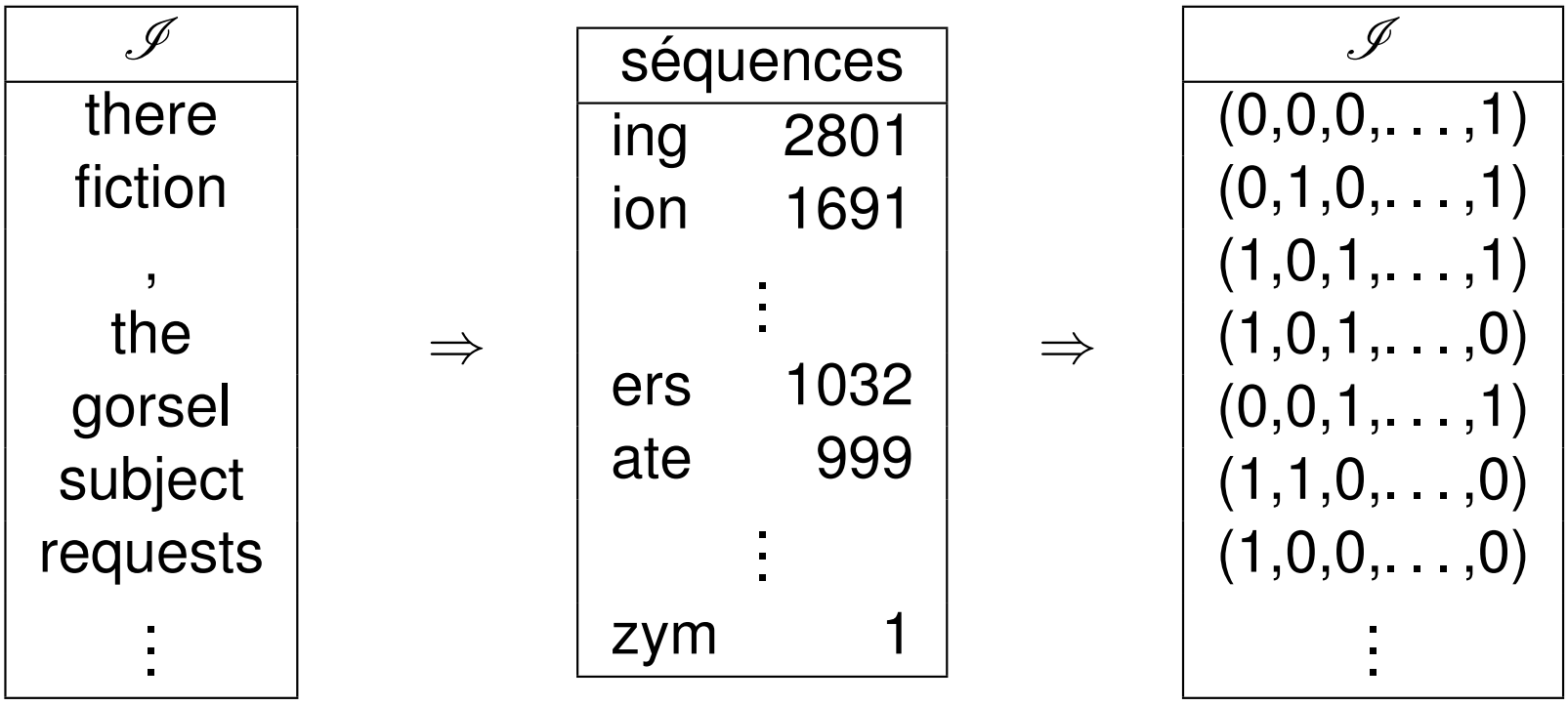
Vector space sampling



Vector space sampling

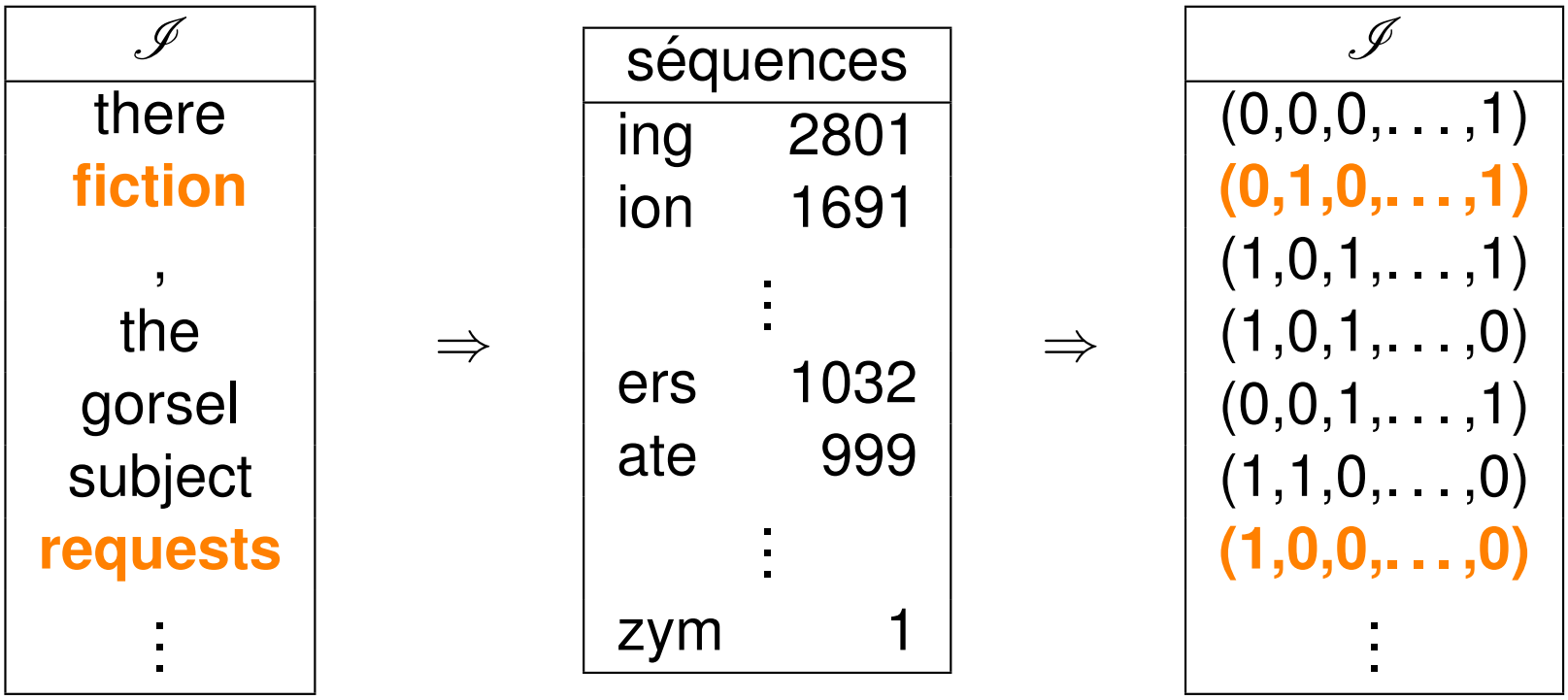


Vector space sampling



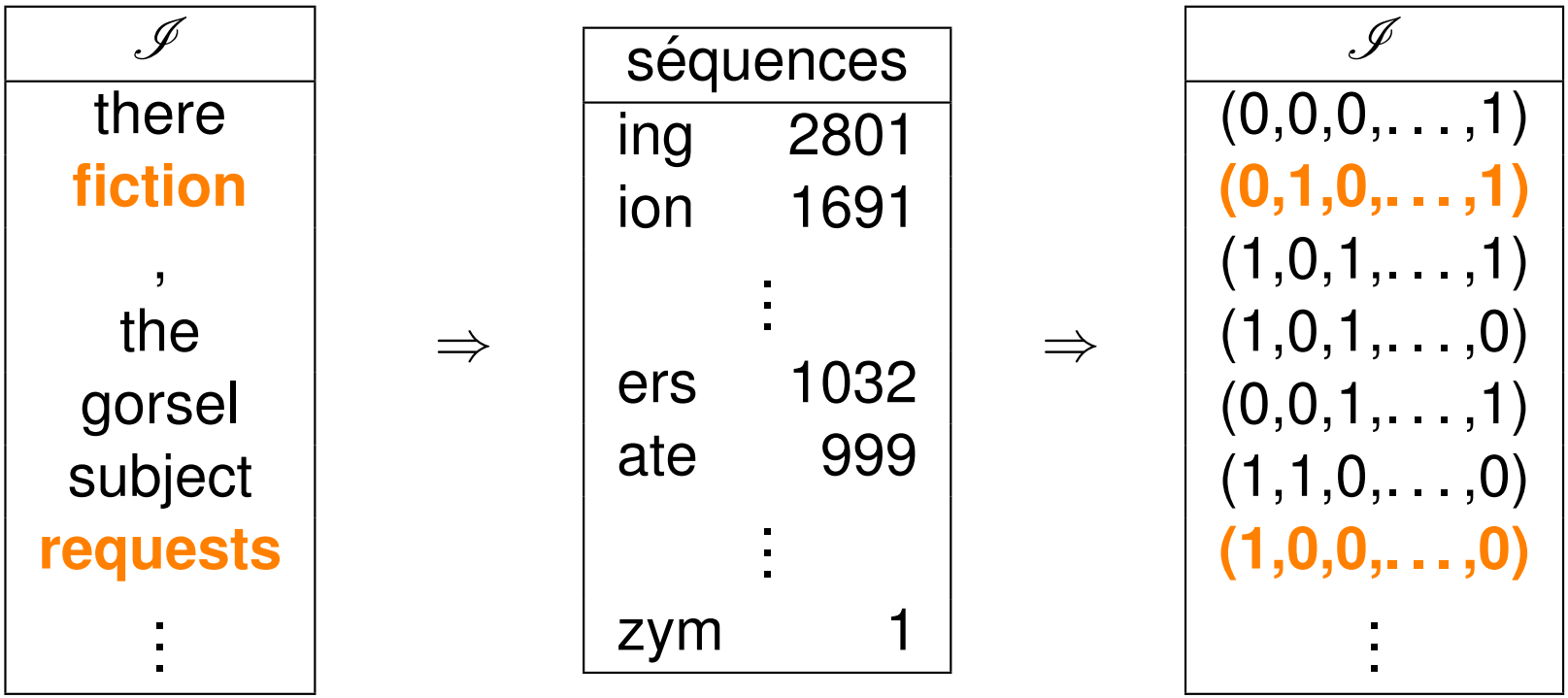
- t ≡ question

Vector space sampling



- t $\equiv$ (0,1,0,...,0)
- similarité cosinus souvent utilisée : similarité(A,B) = cos(θ) = $\frac{A.B}{||A||.||B||} = \frac{\sum_i A_i \times B_i}{\sqrt{\sum_i A_i^2} \times \sqrt{\sum_i B_i^2}}$

Vector space sampling



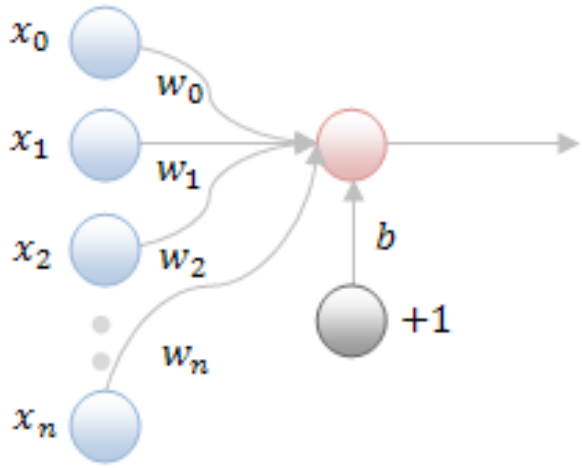
- t $\equiv$ (0,1,0,...,0) $\Rightarrow$ request,requests, query, quest, fiction, ...
- similarité cosinus souvent utilisée : similarité(A,B) = cos(θ) = $\frac{A.B}{||A||.||B||} = \frac{\sum_i A_i \times B_i}{\sqrt{\sum_i A_i^2} \times \sqrt{\sum_i B_i^2}}$

Gérer le bruit (step 3)

- Step 2 peut facilement générer des **millions** de formes candidates . . .
- Plusieurs stratégies proposées:
  - filtrer par la fréquence [Lepage & Denoual, 2005]
  - éliminer les formes non vues dans un grand corpus de formes cibles [Langlais & Patry, 2007]
  - éliminer les formes contenant un n-gramme de caractère non vu à l'entraînement [Lepage, 2007]
  - apprendre à reconnaître les analogies "utiles" [Langlais et al., 2008]

Classification des solutions (bonne/mauvaise)

Perceptron [RosenBlatt, 1958]



- x_j input (valeur réelle)
- w_j poids synaptique (valeur réelle)
- b biais (peut être représenté par un poids supplémentaire et une entrée fixe)
- $h = \sum_j x_j \times w_j + b$ est le signal arrivant au neurone
- l'activation du neurone est déterminée par une fonction d'activation $t \equiv f(h)$ (ex: $f = sign$)

- soit $(x_i, y_i)_{i \in [1,L]}$ un corpus d'entraînement où y_i (ici à valeur dans $\{-1, +1\}$) est la réponse (supervision) et $x_i \in R^{n+1}$
- apprendre les w_j peut se faire en ligne en ajustant itérativement les poids une fois chaque observation x_i rencontrée:
  $\forall j \in [0, n],\ w_j \leftarrow w_j + \eta(y_i - t_i).x_i$
  où η est le learning rate

résout des problèmes linéairement séparables

Classification des solutions (bonne/mauvaise)

Voted-perceptron [Freund:99]

- Entraînement en ligne:

Require: $\{(\mathbf{x}_i, y_i)\}_{i\in[1,L]}$ où $y_i \in \{+1, -1\}$ et $x_i \in R^n$
Ensure: a pool of K perceptrons $\{(\mathbf{v}_k, c_k)\}_{k\in[1,K]}$, $v_k \in R^n$, c_k entier
$k, c_1 \leftarrow 0$
$\mathbf{v}_1 \leftarrow [0...0]^T$
for all epoch **do**
  for all $i \in [1, L]$ **do**
    $\hat{y} \leftarrow sign(\mathbf{v}_k.\mathbf{x}_i)$
    if $\hat{y} \neq y_i$ **then**
      $\mathbf{v}_{k+1} \leftarrow \mathbf{v}_k + y_i\mathbf{x}_i$
      $c_{k+1} \leftarrow 1$
      $k \leftarrow k+1$
    else
      $c_k \leftarrow c_k + 1$

Garanties de convergence (même dans les cas non linéairement séparables)

Classification des solutions (bonne/mauvaise)

Voted-perceptron [Freund:99]

- Test:

$$\hat{y} = sign\left(\sum_k c_k.sign(\mathbf{v}_k.\mathbf{x})\right)$$

- requiert la sauvegarde de K perceptrons et de leur coefficient
- calcul en $O(K \times n)$. On peut également ne retenir qu'un sous-ensemble des perceptrons (à l'extrême un seul, par exemple celui de plus fort c_k).

Traits

- soit une paire d'analogies (source,cible) associant une forme `t` à sa solution `s`:

$$(a \equiv [\texttt{I(x)} : \texttt{I(y)} :: \texttt{I(z)} : \texttt{t}], \hat{a} \equiv [\texttt{O(x)} : \texttt{O(y)} :: \texttt{O(z)} : \texttt{s}])$$

- exemples de traits:
  - degré des analogies source (a) et cible ($\hat{a}$)

[`miracle` : `miraculeux` :: `fable` : `fabuleux`] car:

x	mi	rac	ε	le	ε
y	mi	rac	u	le	ux
z	ε	fab	ε	le	ε
t	ε	fab	u	le	ux

Traits

- soit une paire d'analogies (source,cible) associant une forme `t` à sa solution `s`:

$$(a \equiv [\texttt{I(x)} : \texttt{I(y)} :: \texttt{I(z)} : \texttt{t}], \hat{a} \equiv [\texttt{O(x)} : \texttt{O(y)} :: \texttt{O(z)} : \texttt{s}])$$

- exemples de traits:
  - degré des analogies source (a) et cible ($\hat{a}$)

[`miracle` : `miraculeux` :: `fable` : `fabuleux`] mais aussi:

x	mirac	le
y	mirac	uleux
z	fab	le
t	fab	uleux

degré 2 (nombre minimum de facteurs)

Traits

- soit une paire d'analogies (source,cible) associant une forme `t` à sa solution `s`:

$$(a \equiv [\texttt{I(x)} : \texttt{I(y)} :: \texttt{I(z)} : \texttt{t}], \hat{a} \equiv [\texttt{O(x)} : \texttt{O(y)} :: \texttt{O(z)} : \texttt{s}])$$

- exemples de traits:
  - degré des analogies source (a) et cible ($\hat{a}$)

[`miracle` : `miraculeux` :: `fable` : `fabuleux`] mais aussi:

x	mirac	le
y	mirac	uleux
z	fab	le
t	fab	uleux
cofacteurs	(mirac,fab)	(le,uleux)

Traits

- soit une paire d'analogies (source,cible) associant une forme `t` à sa solution `s`:

$$(a \equiv [\texttt{I(x)} : \texttt{I(y)} :: \texttt{I(z)} : \texttt{t}], \hat{a} \equiv [\texttt{O(x)} : \texttt{O(y)} :: \texttt{O(z)} : \texttt{s}])$$

- exemples de traits:
  - degré des analogies source (a) et cible ($\hat{a}$)

[`miracle` : `miraculeux` :: `fable` : `ueufaxbl`] car

x	mirac	ε	l	ε	ε	ε	ε	e
y	mirac	u	l	eu	ε	x	ε	ε
z	ε	ε	ε	ε	fa	ε	bl	e
t	ε	u	ε	eu	fa	x	bl	ε

Traits

- soit une paire d'analogies (source,cible) associant une forme `t` à sa solution `s`:

$$(a \equiv [\texttt{I(x)} : \texttt{I(y)} :: \texttt{I(z)} : \texttt{t}], \hat{a} \equiv [\texttt{O(x)} : \texttt{O(y)} :: \texttt{O(z)} : \texttt{s}])$$

- exemples de traits:
  - degré des analogies source (a) et cible ($\hat{a}$)

[`miracle` : `miraculeux` :: `fable` : `ueufaxbl`] car

x	mirac	ε	l	ε	ε	ε	ε	e
y	mirac	u	l	eu	ε	x	ε	ε
z	ε	ε	ε	ε	fa	ε	bl	e
t	ε	u	ε	eu	fa	x	bl	ε

  - fréquence de `s`, son rang
  - probabilité minimale, maximale, moyenne (etc.) de `s` selon un modèle n-gramme au niveau des caractères
  - traits sur la présence de cofacteurs,
  - etc.

Apprentissage Analogique

- Analogie
- Principe

Quelques réalisations

Sous le capot

- Définitions de l'analogie (formelle)
- Solveurs d'équations
- Recherche des analogies
- Filtrer le bruit

Applications

- Traduction de mots inconnus
- Traduction de termes
- Translitération

Traduction de mots inconnus

Ressource: corpus WMT'06 [Koehn et Monz, 2006]

UNKS dans le corpus de test (French $\rightarrow$ English):

- 20% de noms propres
- 12% d'expressions numériques (années, pages, etc.)
- 8% de mots composés
- 4% de mots d'emprunt (latin, grec)

$\sim$ **50% de mots *normaux***

Protocole

- $\mathscr{L}_T$: **lexique souche**
  Modèles IBM 2 (s2t ∩ t2s) entraînés sur les T premières paires de phrases de TRAIN.
  - ↪ T: 5 000, 10 000, 100 000, 200 000, and 500 000
- $\mathscr{L}_{ref}$: **lexique de référence**
  Modèles IBM 2 (s2t ∩ t2s) entraînés sur TRAIN.
- Entrées dans un modèle IBM (anglais/roumain, $p > 0.01$):

previous	[1162] *al* (0.155) *anterioare* (0.06113) *precedente* (0.03609) *se* (0.02009) *domnule* (0.01758) *cei* (0.01053) *anterior* (0.01005)
previously	[763] *persoane* (0.0216) *anterior* (0.01674) *puternic* (0.0158) *form* (0.01496) *deputai* (0.01341) *adecvat* (0.01315) *drepturilor* (0.01305) *sau* (0.01186) *la* (0.01115) *lor* (0.01087) *zona* (0.01042) *de-a* (0.01037) *suplimentare* (0.01003)

Protocole

- Traduire les mots de TEST ne contenant pas de chiffre et présents dans $\mathscr{L}_{ref}$, **mais** inconnus de $\mathscr{L}_T$.

- Deux mesures:

rappel	% de mots UNKS avec une traduction correcte selon $\mathscr{L}_{ref}$
précision	% nombre de mots avec une traduction correcte selon $\mathscr{L}_{ref}$

Approches de base (baselines)

X $\equiv$ mot inconnu

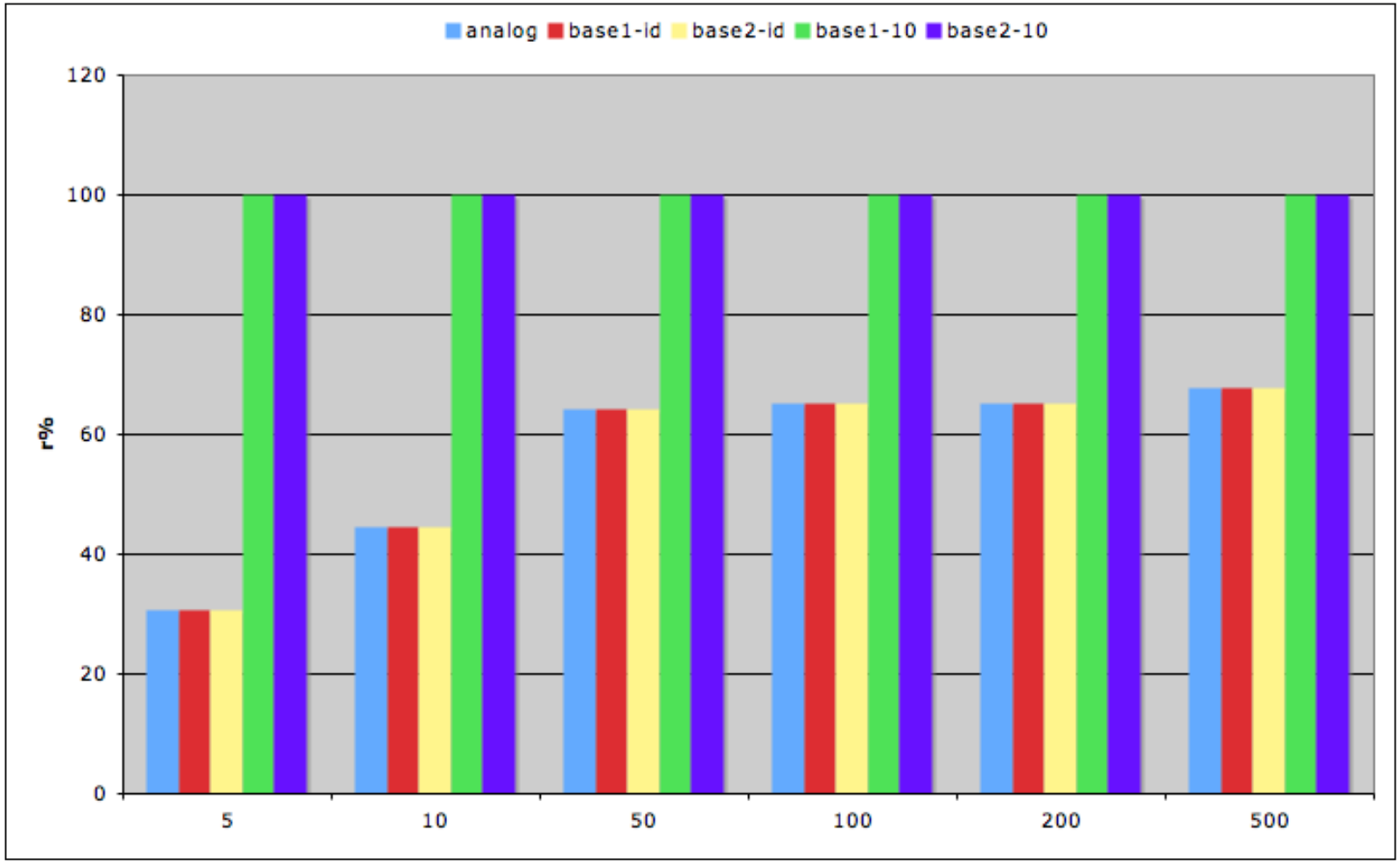
B1 $\hat{T} = \text{argmin}_{T \in \mathscr{O}} \text{edit-dist}(T, X)$

signalaient $\rightarrow$ *signalling*, *signalled*, *salient*, ...

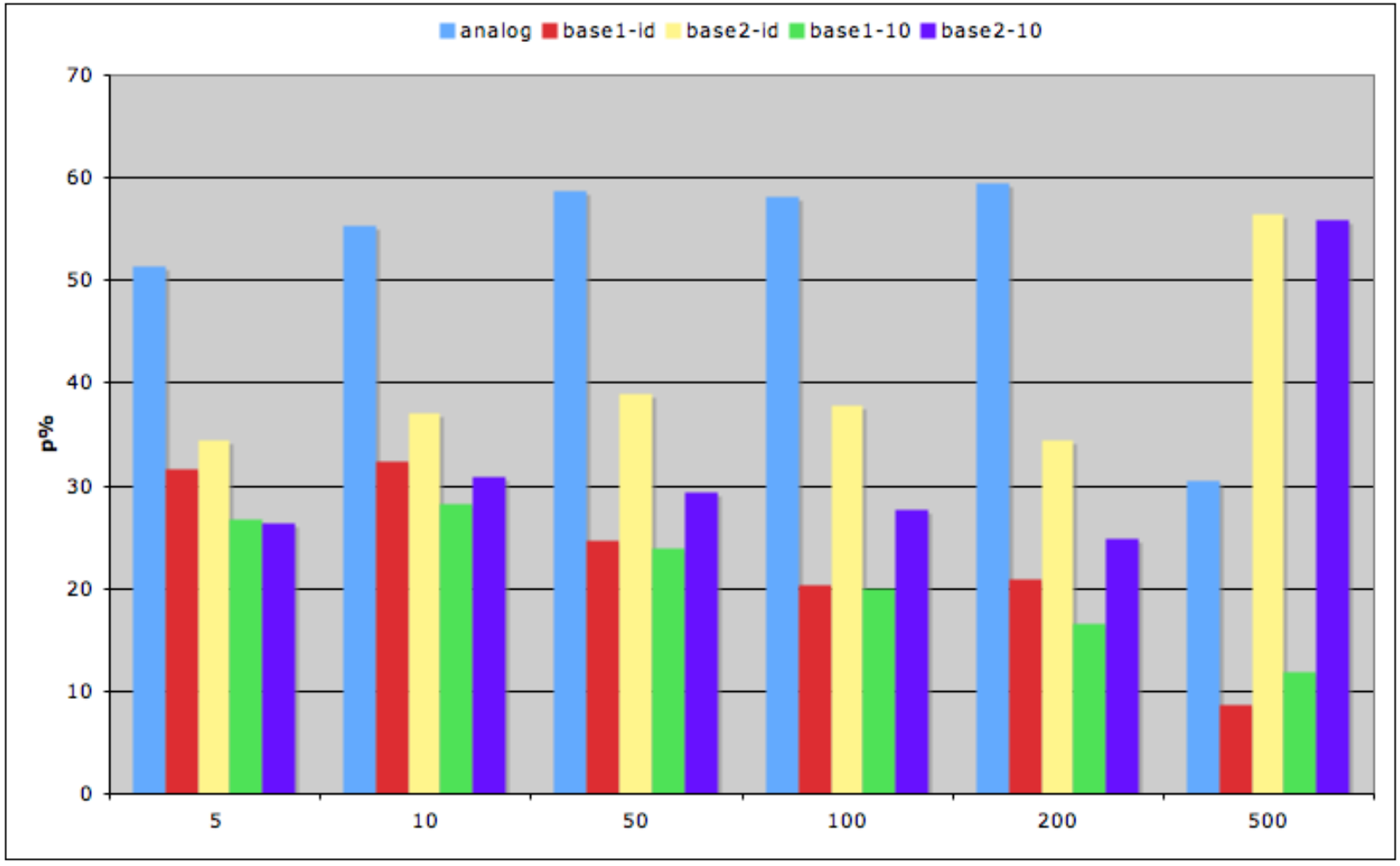
B2 $\hat{T} = \text{argmax}_{S \in V(X)} \mathit{proj}_{\mathscr{L}}(S)$ où
$V(X) = \text{argmin}_{S \in \mathscr{I}} \text{edit-dist}(X, S)$

signalaient $\rightarrow$ *signalement*, *signeraient*, *signalent*, ...
$\rightarrow$ *indicates*, *sign*, *signalling*, ...

Recall (French→English)



Précision (French→English)



Cas de $T = 500\,000$ (French$\rightarrow$English)

- Only a few words seen in $\mathscr{L}_{ref}$, but not in $\mathscr{L}_{500\,000}$
  - 34 words (in-domain), 1.2 translations sanctioned by $\mathscr{L}_{ref}$
  - $\sim$ 1/3 of the reference translations are wrong

- exemples:
  - **perquisitions**:

A	*seizures* (76) *seizure* (17) *raid* (1)
$\mathscr{L}_{ref}$	house-to-house

  - **non-discriminatoire**:

A	*non-discriminatory* (72) *non-discrimination* (47) *nondiscrimination* (24) *nondiscriminatory* (23) *outlawing* (20) *race* (20) *prosperity* (20) *antidiscrimination* (20) *anti-discrimination* (18) *discrimination* (17)
$\mathscr{L}_{ref}$	affirmative

Évaluation manuelle

contournant	(*49 candidats*)
A ◇ **(circumventing,55)** (undermining,20) (evading,19) (circumvented,17) (overturning,16) (circumvent,15) (circumvention,15) (bypass,13) (evade,13) (skirt,12) . . .	
$\mathscr{L}_{ref}$ ◇ **skirting**, **bypassing**, by-pass, overcoming	

- 80% des **mots ordinaires** recoivent une traduction valide en première position par analogie
- 35% dans le cas de B2

Impact en traduction (vers l'anglais)

- Ajout dans le modèle de traduction de la première traduction produite par analogie pour chaque mot inconnu (lexique souche = $\mathscr{L}_{ref}$)
- Évalué sur les phrases contenant **au moins** un mot inconnu

	French		Spanish		German	
	WER	BLEU	WER	BLEU	WER	BLEU
base	61.8	22.74	54.0	27.00	69.9	18.15
+B2	61.8	22.72	54.2	26.89	70.3	18.05
+A	**61.6**	**22.90**	**53.7**	**27.27**	**69.7**	**18.30**
sentences	387		452		814	

↪ améliorations stables mais non-significatives

Traduction de segments inconnus

French → English

- $\mathscr{L}_{ref}, \mathscr{L}_T$: phrase table

expulsent ◇ *expelling* (36) *expel* (31) *are expelling* (23) *are expel* (10)
focaliserai ◇ *focus* (10) *focus solely* (9) *concentrate all* (9) *(will focus* (9) *will placing* (9)
dépasseront ◇ *will exceed* (4) *(exceed* (3) *will be exceed* (3) *we go beyond* (2) *will be exceeding* (2)
non-réussite *de* ◇ *lack of success for* (4) *lack of success of* (4) *lack of success* (4)
que vous ***subissez*** ◇ *you are experiencing* (2)

- précision: 55%
- rappel: 10% (trop de filtrage pendant step-1)

Enrichissement d'une *phrase table*

- TRAIN: table de segments entraînée sur WMT'06 (fr2en)

```
comprehensive safety ||| sécurité globale ||| 0.111111 0.000569718
comprehensive security ||| sécurité globale ||| 0.222222 0.000789918
general safety ||| sécurité globale ||| 0.111111 0.000161561
global safety ||| sécurité globale ||| 0.111111 0.00126761
global security ||| sécurité globale ||| 0.444444 0.00175755
advocate a modern comprehensive security ||| favorables à un concept de sécurité globale moderne ||
```

- Sélecteur (step-3): aucun

Exemples

- âgées à leur sort. [1079]
  old to die . (57) *old on their own .* (56) *old in the lurch .* (53) *old to their fate .* (41) *very old to die .* (35) . . .
- ’ acquis soient transposées [3610]
  acquis are transposed (47) *acquis be transposed with* (38) *acquis will be transposed* (37) . . .
- a caractérisé la réunification allemande [3655]
  has characterised of german reunification, (24) *has characterised german reunification* (20) . . .
- acceptables , sans mettre en [9985]
  acceptable without calling into (23) *acceptable, without calling into* (21) . . .
- a été discutée et [406223]
  were debated and (151) *was discussed this and* (133) *was discusseds thi and* (123) *has been discussed and has* (119). . .

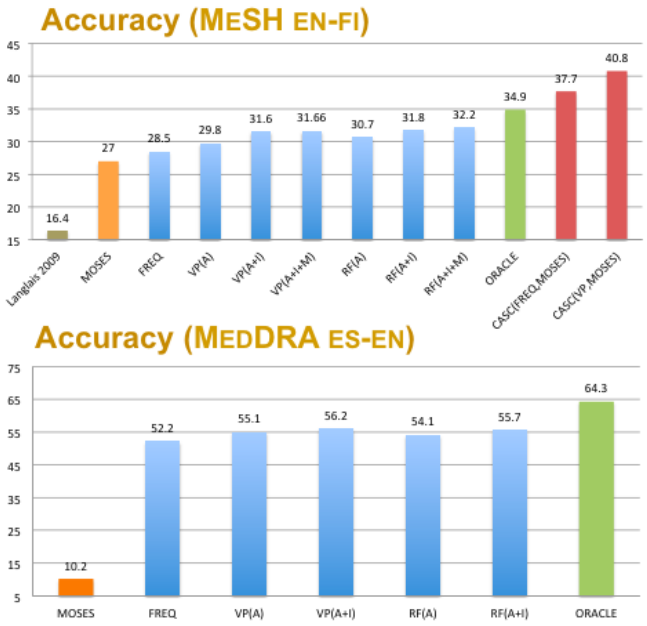
Traduction de termes médicaux

- but: comparer l'apprentissage analogique à la traduction statistique par segments (Koehn et al., 2003) où l'unité de base est le caractère (sinon c'est trop mauvais)
- paires de langues variées: RU-EN, FI-EN, SW-EN, ES-EN, FR-EN
- un besoin réel

Corpora

	TRAIN			TEST	DEV
	NB	AVG. LENGTH	NB	OOV	OOV
MESH (Medical Subject Headings thesaurus)					
FI	19787	19.3	1000	65.0	63.8
FR	17230	21.5	1000	35.8	36.8
RU	21407	38.5	1000	42.3	45.1
ES	19201	21.5	1000	37.4	34.9
SW	17090	17.3	1000	69.3	70.0
MEDDRA (Medical Drug Regulatory Activities thesaurus)					
ES	65276	34.6	1000	7.1	7.1

Résultats



- ATRANS ≥ PB-SMT
  - ✓ Similar tendencies for other directions
  - ✓ Cascading ATRANS and PB-SMT improves drastically over each system
- Rescoring (RF) at a small advantage

- ATRANS >> Moses
  - ➲ Similar observation for EN-ES
  - ➲ ATRANS (much) less silent
- Rescoring (RF) > classification (VP)

Exemples

SW	aikakauslehdet aiheena
REF	periodicals as topic
ANA	periodicals as topic
SMT	timenancylages, topic

SW	alfasalpaajat
REF	adrenergic alpha-antagonists
ANA	adrenergic alpha-antagonists
SMT	alphablockers

SP	instituciones de atención ambulatoria
REF	ambulatory care facilities
ANA	ambulatory care facilities
SMT	institutions, atention ambulatory

FI	märkivä kilpirauhastulehdus
REF	thyroiditis, suppurative
ANA	thyroiditis suppurativa
SMT	rativa thyroid glandorum

FR	malformations de la machoîre
REF	jaw abnormalities
ANA	jaw congenital abnormalities
SMT	malformations jawory

FI	rasva-alkoholit
REF	fatty alcohols
ANA	lipid alcohols
SMT	fatty-alcohols

Translitération

NEWS 2009 EN-CH [NEWS09]

	TRAIN	DEV	TEST
examples	31 961	2 896	2 896
symbols	26	26	26
symbols	370	275	283

examples	
Emission	埃米申
Blagrove	布格夫
Aposhian	阿波希安

NEWS 2009 evaluation script:

- **ACC: accuracy of the first solution**
- F_1: partial credit proportional to the longest subsequence between the 1st candidate and the reference
- MRR: Mean Reciprocal Rank

System Tested

- Pure analogical devices:
  - ana-freq a_1
    solutions ranked by frequency
  - ana-svm$_A$ a_2
    solutions ranked by an SVM trained on analogical-based features
- Moses m
  default setting (word $\equiv$ character) ?
- Hybrids:
  - ana-svm$_{A+M}$ am_1
    solutions ranked by an SVM trained on analogical **and** moses-based features
  - am_2 : cascading am_1 and m
    trust AM_1 whenever a solution is produced, m otherwise

Results

Configuration		ACC	F_1	MRR	*rank*
a_1	ana-freq	**56.6**	79.1	63.0	16
a_2	ana-svm$_A$	**58.0**	80.0	58.8	15
m	moses	**66.6**	85.9	66.6	6
am_1	ana-svm$_{A+M}$	**63.4**	82.0	64.1	10
am_2	casc(am_1,m)	**68.5**	86.9	69.0	4
	last NEWS 2009	**19.9**	60.6	22.9	23
	first NEWS 2009	**73.1**	89.5	81.2	1

1- None of the systems tested outperform the best system at NEWS 2009

Results

Configuration		ACC	F_1	MRR	*rank*
a_1	ana-freq	**56.6**	79.1	63.0	16
a_2	ana-svm$_A$	**58.0**	80.0	58.8	15
m	moses	**66.6**	85.9	66.6	6
am_1	ana-svm$_{A+M}$	**63.4**	82.0	64.1	10
am_2	casc(am_1,m)	**68.5**	86.9	69.0	4
	last NEWS 2009	**19.9**	60.6	22.9	23
	first NEWS 2009	**73.1**	89.5	81.2	1

2- The best system we tested would have ranked 4th

Results

Configuration		ACC	F_1	MRR	*rank*
a_1	ana-freq	**56.6**	79.1	63.0	16
a_2	ana-svm_A	**58.0**	80.0	58.8	15
m	moses	**66.6**	85.9	66.6	6
am_1	ana-svm_{A+M}	**63.4**	82.0	64.1	10
am_2	casc(am_1,m)	**68.5**	86.9	69.0	4
	last NEWS 2009	**19.9**	60.6	22.9	23
	first NEWS 2009	**73.1**	89.5	81.2	1

3- Learning to recognize good solutions is preferable to ranking by frequency

Results

Configuration		ACC	F_1	MRR	*rank*
a_1	ana-freq	**56.6**	79.1	63.0	16
a_2	ana-svm_A	**58.0**	80.0	58.8	15
m	moses	**66.6**	85.9	66.6	6
am_1	ana-svm_{A+M}	**63.4**	82.0	64.1	10
am_2	casc(am_1,m)	**68.5**	86.9	69.0	4
	last NEWS 2009	**19.9**	60.6	22.9	23
	first NEWS 2009	**73.1**	89.5	81.2	1

4- Moses outperforms pure analogical devices
(silence rate of analogical devices: 3.7%)

Results

Configuration		ACC	F_1	MRR	*rank*
a_1	ana-freq	**56.6**	79.1	63.0	16
a_2	ana-svm$_A$	**58.0**	80.0	58.8	15
m	moses	**66.6**	85.9	66.6	6
am_1	ana-svm$_{A+M}$	**63.4**	82.0	64.1	10
am_2	casc(am_1,m)	**68.5**	86.9	69.0	4
	last NEWS 2009	**19.9**	60.6	22.9	23
	first NEWS 2009	**73.1**	89.5	81.2	1

5- Adding a few Moses-based features helps classification a lot

Results

Configuration		ACC	F_1	MRR	*rank*
a_1	ana-freq	**56.6**	79.1	63.0	16
a_2	ana-svm_A	**58.0**	80.0	58.8	15
m	moses	**66.6**	85.9	66.6	6
am_1	ana-svm_{A+M}	**63.4**	82.0	64.1	10
am_2	casc(am_1,m)	**68.5**	86.9	69.0	4
	last NEWS 2009	**19.9**	60.6	22.9	23
	first NEWS 2009	**73.1**	89.5	81.2	1

6- Higher precision of the analogical device over Moses