Probabilistic Finite State Machines

Using and building

Franck Thollard

`http://eurise.univ-st-etienne.fr/∼thollard`

EURISE

Saint-Étienne

France

Contents

- Probabilistic Finite State Machines

Contents

- Probabilistic Finite State Machines
- What learning means

Contents

- Probabilistic Finite State Machines
- What learning means
- Learnability results

Saint-Etienne

Contents

- Probabilistic Finite State Machines
- What learning means
- Learnability results
- Algorithmic issues

Saint-Etienne

Contents

- Probabilistic Finite State Machines
- What learning means
- Learnability results
- Algorithmic issues
- Experimental issues

Contents

- Probabilistic Finite State Machines
  - n-grams / MM / PST / A-PDFA
  - PDFA / Residual automata
  - Other : PFA / HMM, PCFG, ...
  - Choosing a model
- What learning means
- Learnability results
- Algorithmic issues
- Conclusions

Saint-Etienne

The probability of a sequence

Computation using the chain rule:

$$\begin{aligned} P(he\ reads\ a\ book) &= P(he) \times P(reads|he) \\ &\times P(a|he\ reads) \times P(book|he\ reads\ a) \end{aligned}$$

The probability of a sequence

Computation using the chain rule:

$$
\begin{aligned}
P(he\ reads\ a\ book) \quad &= \quad P(he) \times P(reads|he) \\
&\quad \times P(a|he\ reads) \times P(book|he\ reads\ a)
\end{aligned}
$$

and more generally:

$$P(w_1 w_2 \dots w_n) = P(w_1) \times P(w_2|w_1) \times \dots \times P(w_n|w_1 w_2 \dots w_{n-1})$$

The probability of a sequence

Computation using the chain rule:

$$\begin{aligned} P(he\ reads\ a\ book) &= P(he) \times P(reads|he) \\ &\times P(a|he\ reads) \times P(book|he\ reads\ a) \end{aligned}$$

and more generally:

$$P(w_1 w_2 \ldots w_n) = P(w_1) \times P(w_2|w_1) \times \ldots \times P(w_n|w_1 w_2 \ldots w_{n-1})$$

Definition: $w_1 w_2 \ldots w_{n-1}$ is called the **history**.

Saint-Etienne

Some models

See (Vidal *et al.*, 2005) for a survey

- n-grams / MM $\Leftrightarrow$ k-testables automata

Saint-Etienne

Some models

See (Vidal *et al.*, 2005) for a survey

- n-grams / MM $\Leftrightarrow$ k-testables automata
- Probabilistic Automata without cycles

Saint-Etienne

Some models

See (Vidal *et al.*, 2005) for a survey

- n-grams / MM $\Leftrightarrow$ k-testables automata
- Probabilistic Automata without cycles
- Probabilistic Deterministic Automata

Some models

See (Vidal *et al.*, 2005) for a survey

- n-grams / MM $\Leftrightarrow$ k-testables automata
- Probabilistic Automata without cycles
- Probabilistic Deterministic Automata
- Residual Automata (Esposito & al., 2002)

Saint-Etienne

Some models

See (Vidal *et al.*, 2005) for a survey

- n-grams / MM $\Leftrightarrow$ k-testables automata
- Probabilistic Automata without cycles
- Probabilistic Deterministic Automata
- Residual Automata (Esposito & al., 2002)
- Probabilistic non-deterministic Automata ($\Leftrightarrow$ HMM)

Some models

See (Vidal *et al.*, 2005) for a survey

- n-grams / MM $\Leftrightarrow$ k-testables automata
- Probabilistic Automata without cycles
- Probabilistic Deterministic Automata
- Residual Automata (Esposito & al., 2002)
- Probabilistic non-deterministic Automata ($\Leftrightarrow$ HMM)
- PCFG: Probabilistic Context-Free Grammars

Some models

See (Vidal *et al.*, 2005) for a survey

- n-grams / MM $\Leftrightarrow$ k-testables automata
- Probabilistic Automata without cycles
- Probabilistic Deterministic Automata
- Residual Automata (Esposito & al., 2002)
- Probabilistic non-deterministic Automata ($\Leftrightarrow$ HMM)
- PCFG: Probabilistic Context-Free Grammars

Note: models define a pdf on Σ^n, for each n
add of eos symbol $\Rightarrow$ pdf on Σ^*

Saint-Etienne

The n-grams model (1/3)

Assumption: the history is supposed bound

Saint-Etienne

The n-grams model (1/3)

Assumption: the history is supposed bound

Example: history of size one $\Rightarrow$ 2-grams (known as `bigram`)

The n-grams model (1/3)

Assumption: the history is supposed bound

Example: history of size one $\Rightarrow$ 2-grams (known as `bigram`)

$$
\begin{aligned}
P(he\ reads\ a\ book) &= P(he) \times P(reads|he) \\
&\quad \times P(a|reads) \times P(book|a)
\end{aligned}
$$

The n-grams model (2/3)

Estimating n-grams probabilities

The probabilities are estimated using a corpus by counting occurrences of the n-uplets :

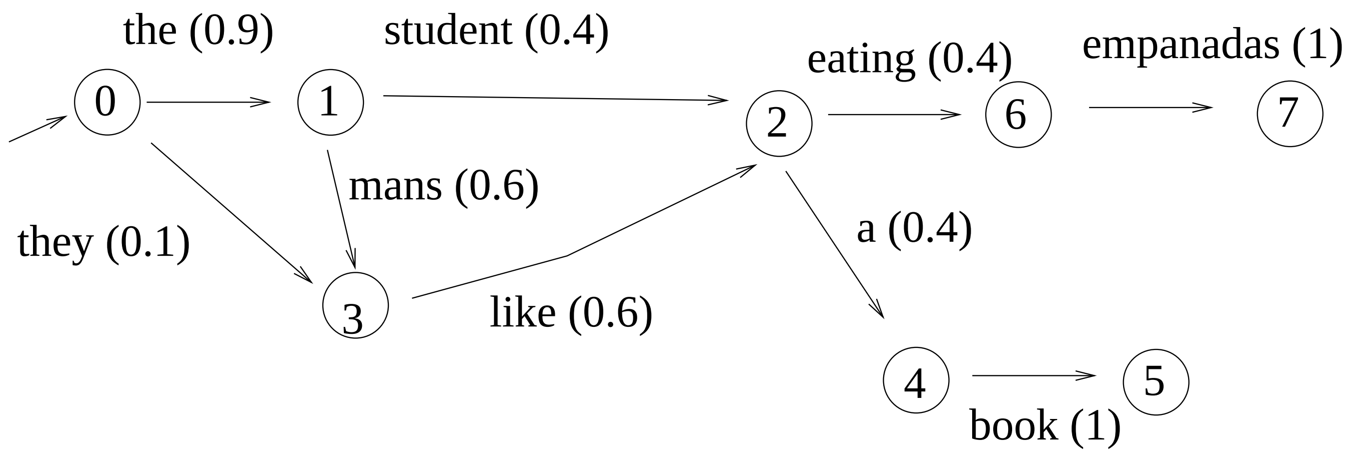
$$P(book|a) = \frac{\mathrm{Ct}(a\ book)}{\mathrm{Ct}(a)}$$

smoothing

Saint-Etienne

The n-grams (3/3)

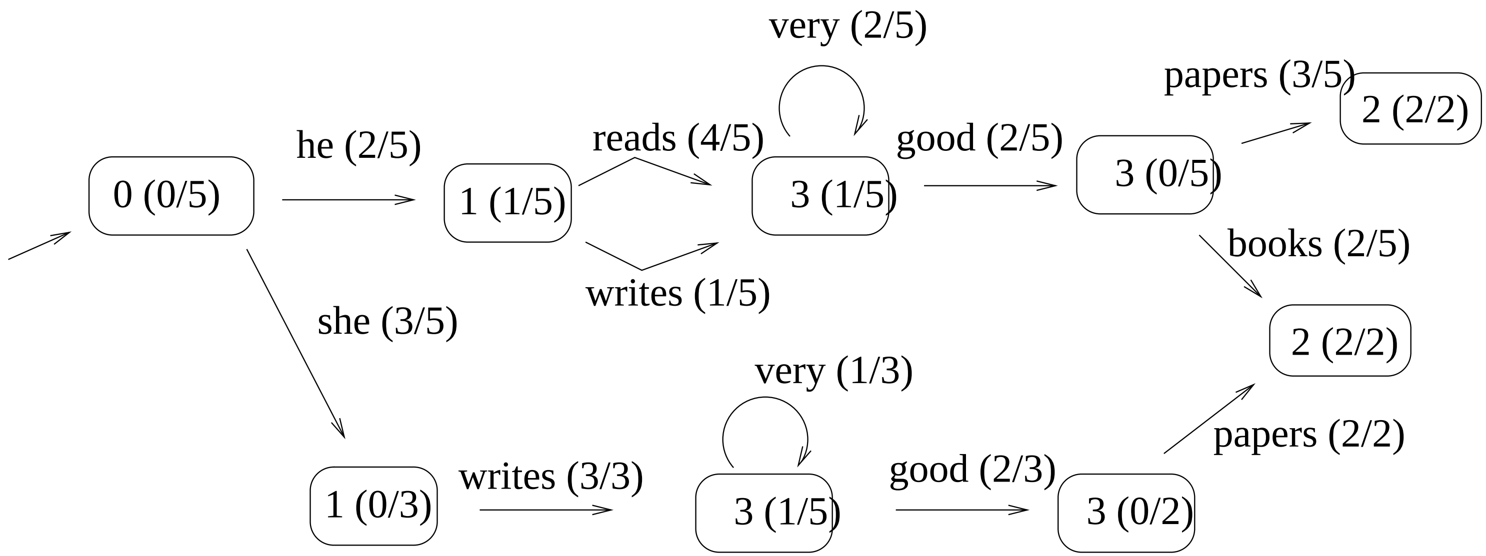
Automata representation of n-grams



Cyclic Automaton

unbounded history

A Probabilistic automaton



Note: $P(papers|\ldots)$ depends on an unbound history.

Choosing a model

Model	Good	Bad
n-gram	Easy to build, Good estimates	Assumption can be false Big model

Saint-Etienne

Choosing a model

Model	Good	Bad
n-gram	Easy to build, Good estimates	Assumption can be false Big model
A-PDFA	History larger	Building the structure

Saint-Etienne

Choosing a model

Model	Good	Bad
n-gram	Easy to build, Good estimates	Assumption can be false Big model
A-PDFA	History larger	Building the structure
PDFA	Unbound history Parsing time Correct results Size of the model	Building the structure

Saint-Etienne

Choosing a model

Model	Good	Bad
n-gram	Easy to build, Good estimates	Assumption can be false Big model
A-PDFA	History larger	Building the structure
PDFA	Unbound history Parsing time Correct results Size of the model	Building the structure
PFA / HMM	Unbound history Hand made structure Size of the model	Parsing time or "Loose" of proba in non determinism

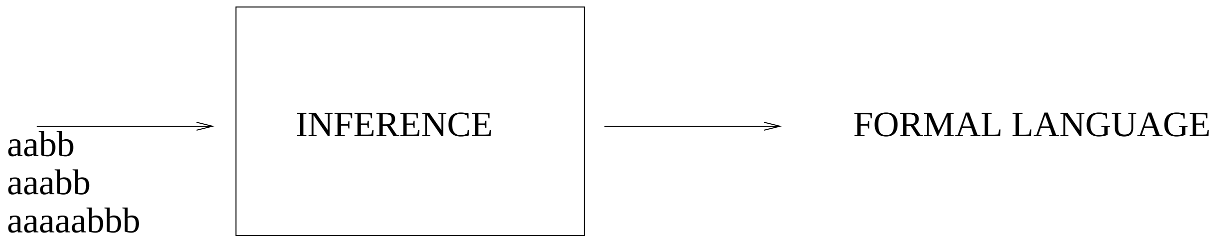
Saint-Etienne

Contents

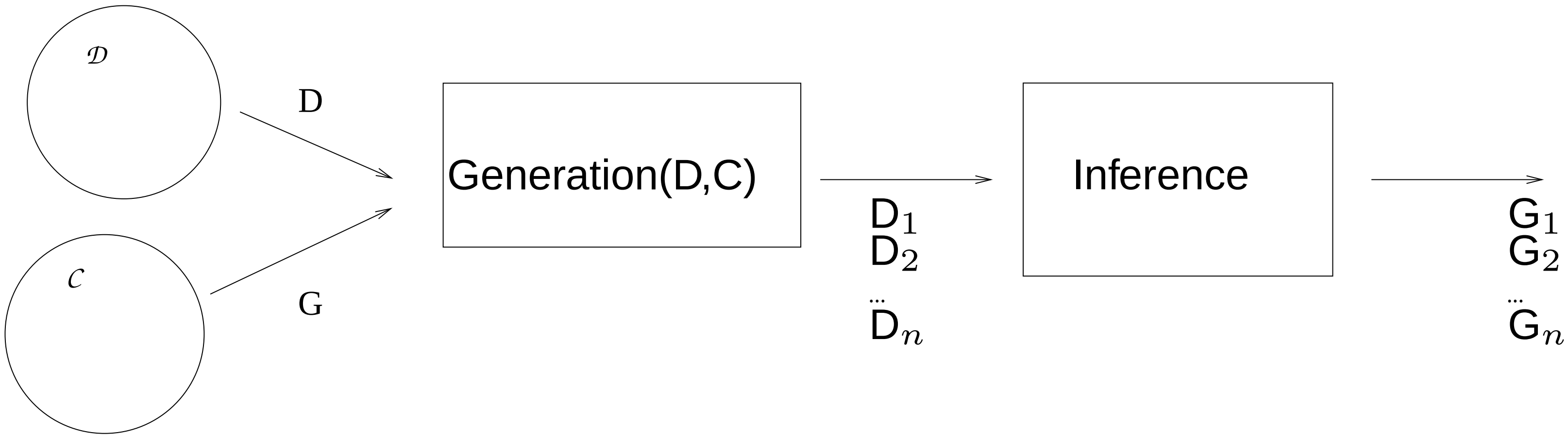
- Probabilistic Finite State Machines
- What learning means
  - What learning means ?
  - What can be learned ?
- Learnability results
- Algorithmic issues
- Experimental issues
- Conclusions

Machine Learning Assumption

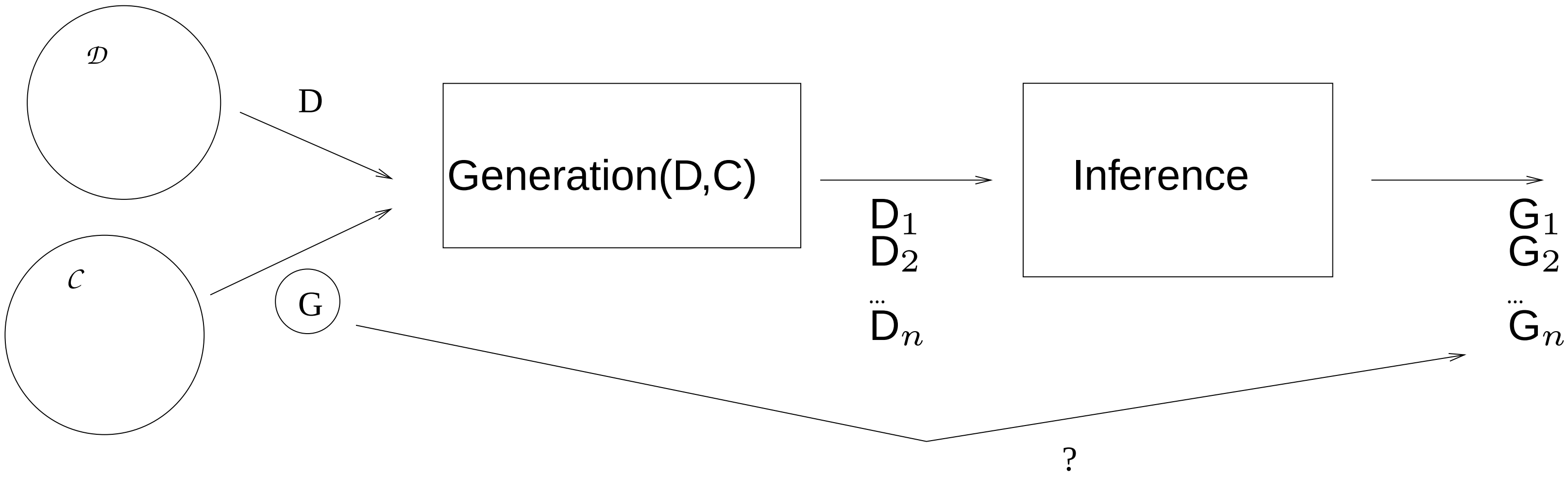
Formal Language Learning



What learning means ?



What learning means ?



What learning means ?

Formalization

- Learning criterion

What learning means ?

Formalization

- Learning criterion
- Learnability results (w.r.t. automata)

Saint-Etienne

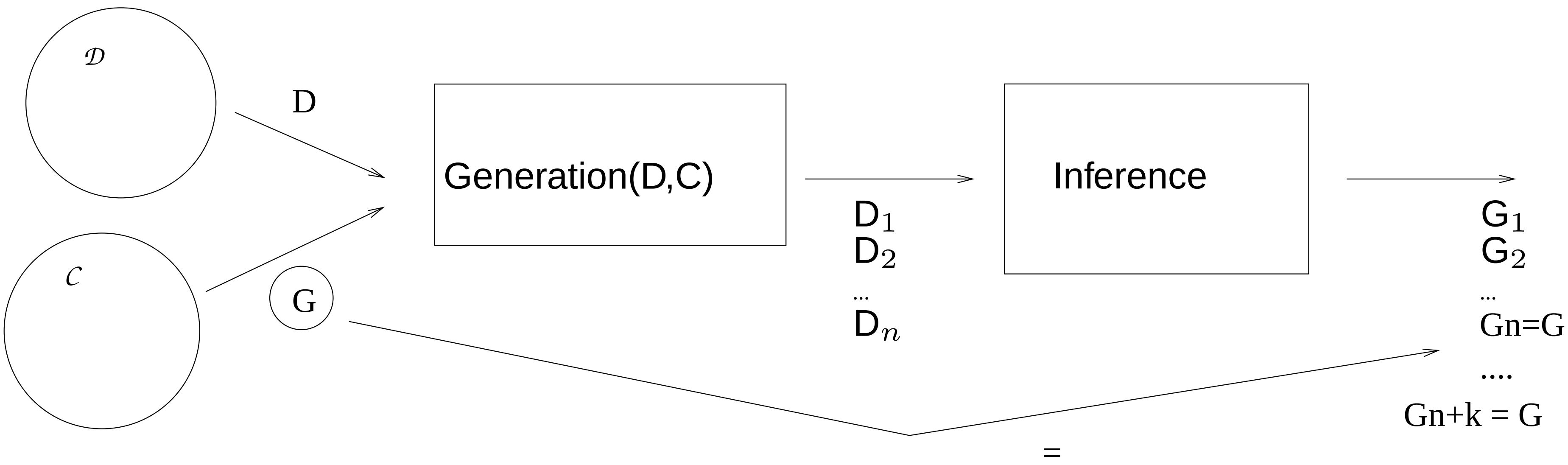
What learning means ?

Formalization

- Learning criterion
  - Identification in the limit
  - PAC learning

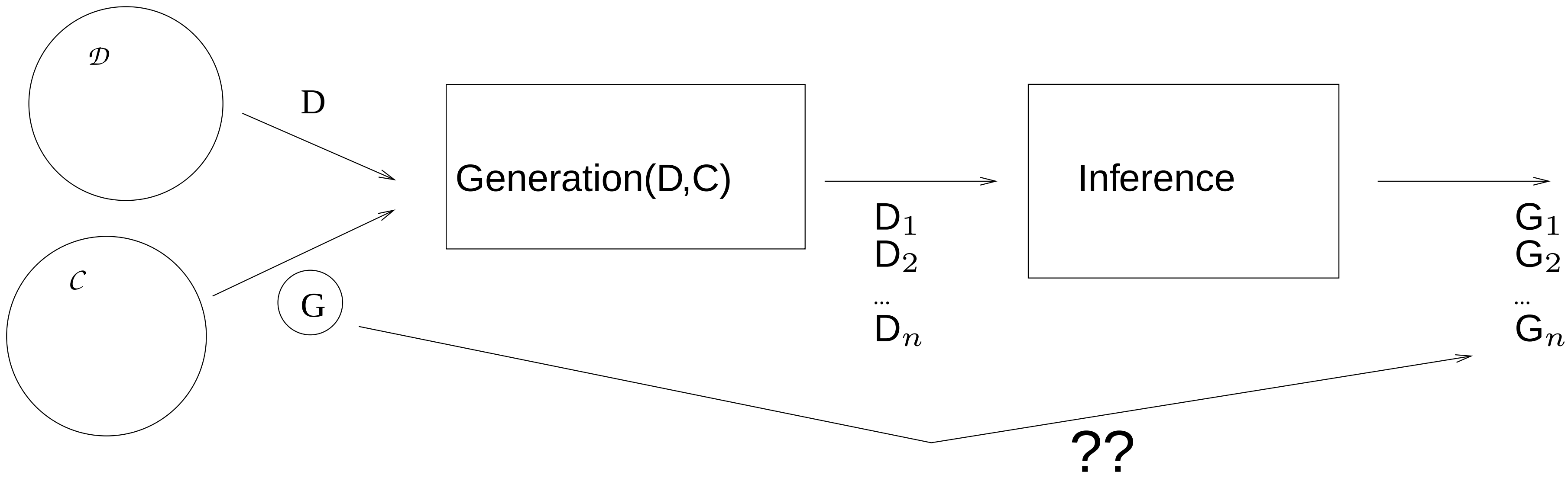
Saint-Etienne

Identification in the limit



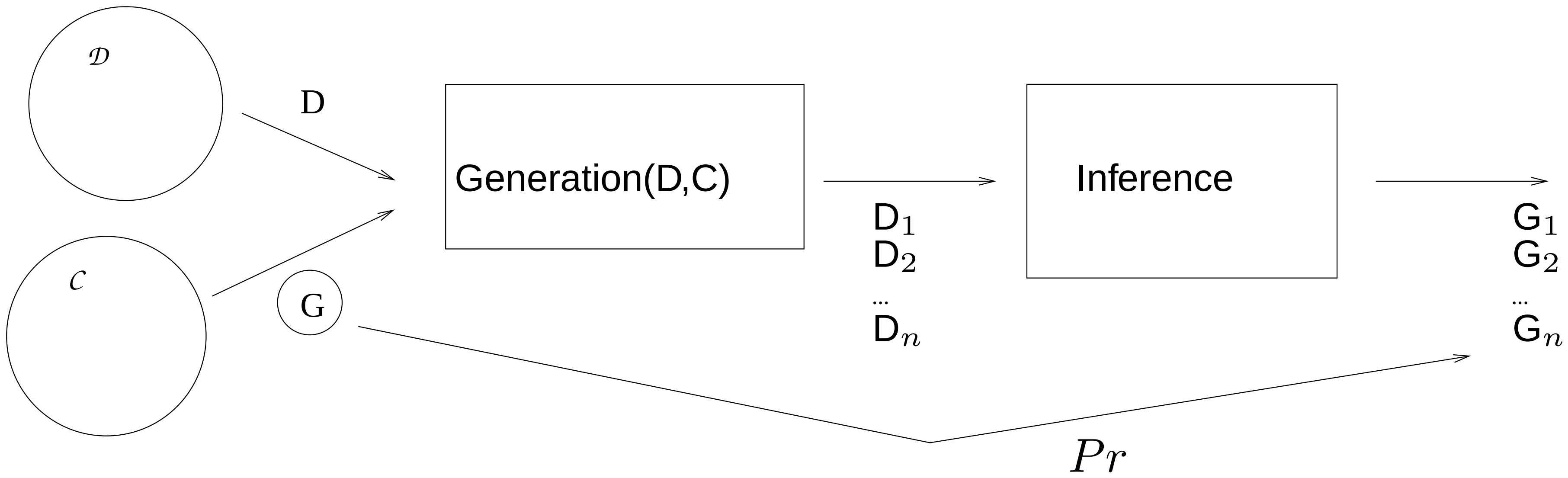
PAC Learning

Proba–D–PAC



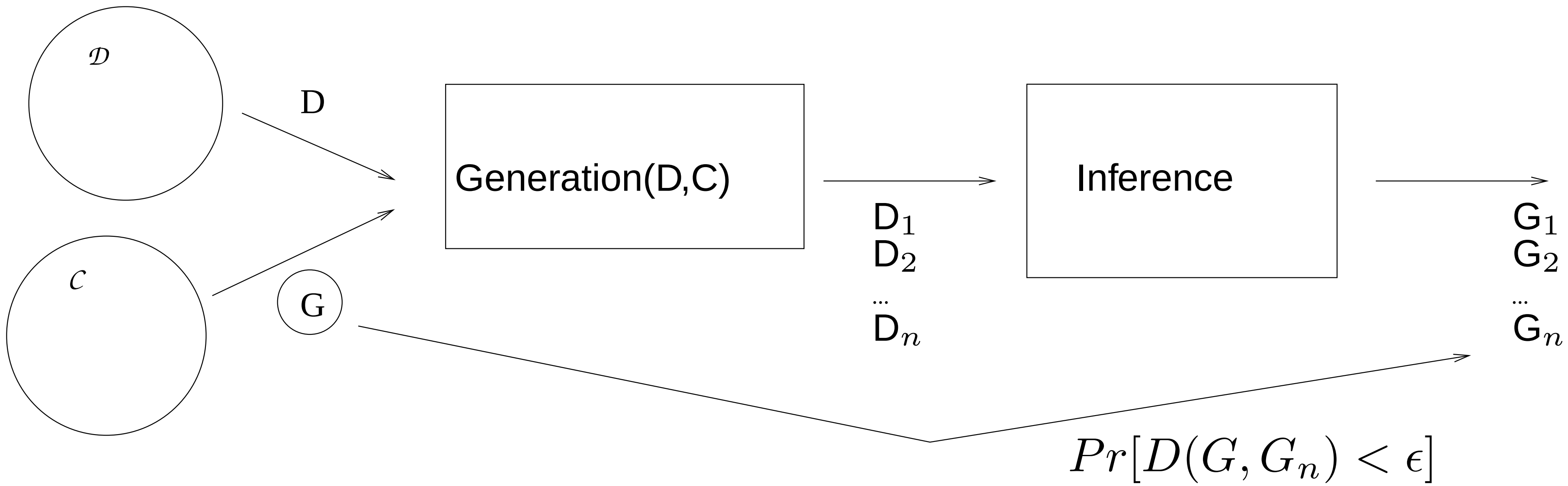
PAC Learning

Proba–D–PAC



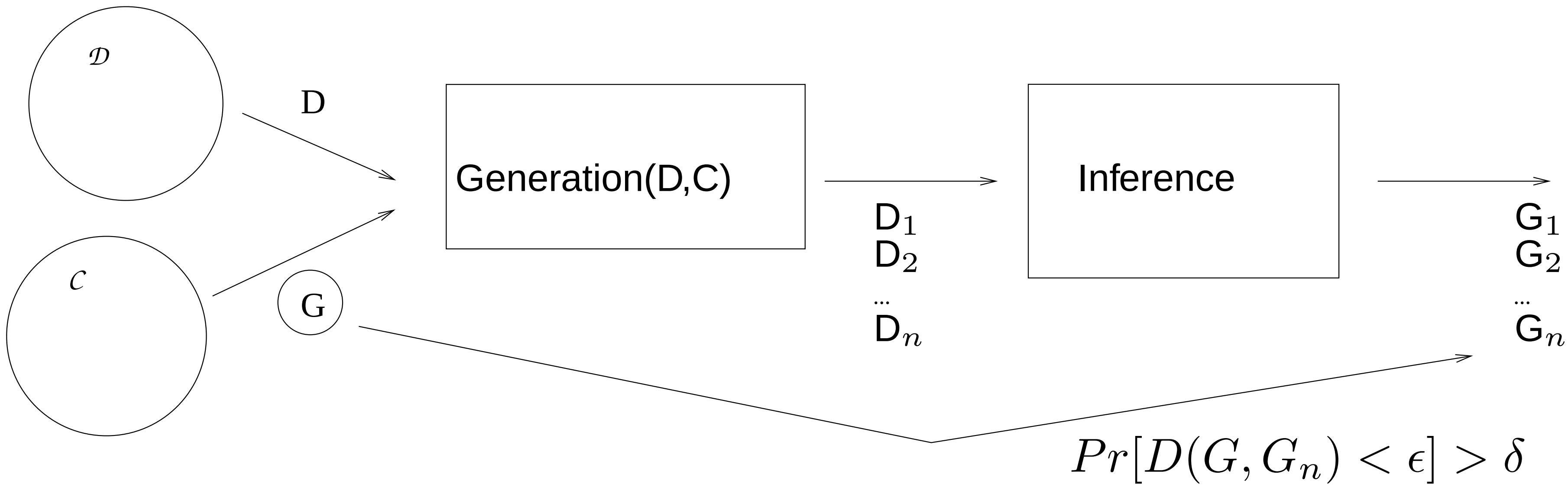
PAC Learning

Proba–D–PAC



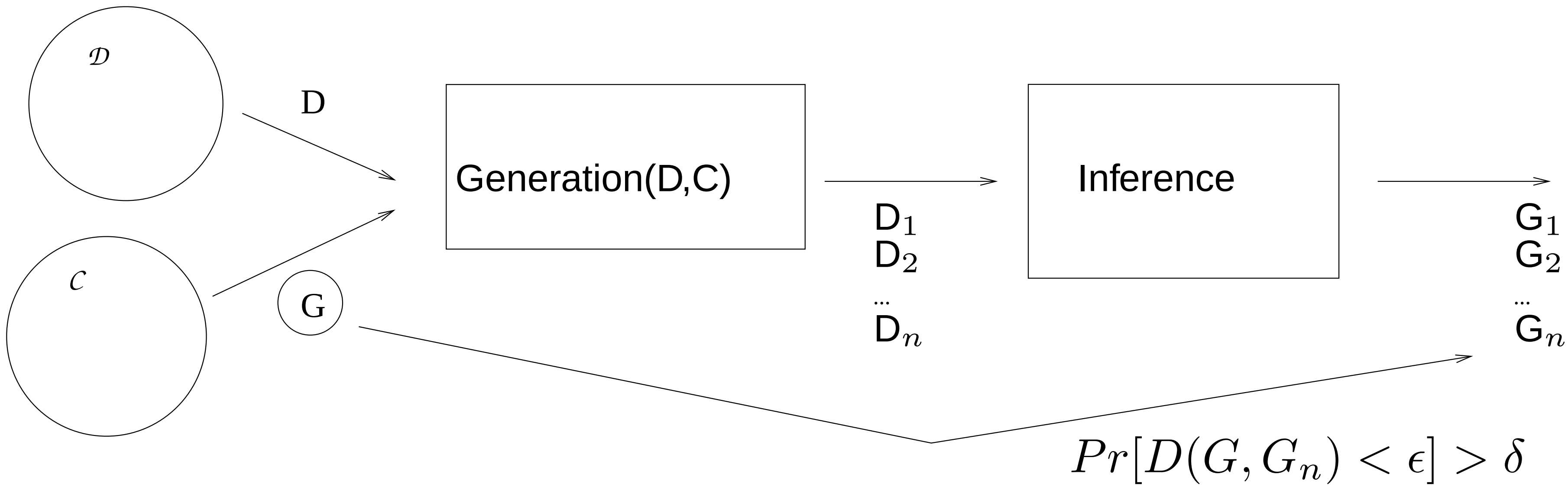
PAC Learning

Proba–D–PAC



PAC Learning

Proba–D–PAC



Link between

Precision / Confidence / number of examples / complexity of the class / . . .

Identification in the limit

(With Proba One)

- The class of **recursively enumerable** languages can be identified in the limit with probability one (Horning, 1969).

Saint-Etienne

Identification in the limit

(With Proba One)

- The class of **recursively enumerable** languages can be identified in the limit with probability one (Horning, 1969).
- The class of **probabilistic automata** can be identified in the limit with probability one (**constructive proof**) (Carrasco, 1999).

Saint-Etienne

Proba-Poly-PAC (1/2)

Possible

- Proba-d_∞-PAC PFA (Angluin 88, lemma 14)

Proba-Poly-PAC (1/2)

Possible

- Proba-d_∞-PAC PFA (Angluin 88, lemma 14)
- Proba-KL-PAC Unigram with unknown vocabulary (Mc Allester & Shapire, 2000)

Saint-Etienne

Proba-Poly-PAC (1/2)

Possible

- Proba-d_∞-PAC PFA (Angluin 88, lemma 14)
- Proba-KL-PAC Unigram with unknown vocabulary (Mc Allester & Shapire, 2000)
- Proba-KL-PAC PFA on Σ^n, Σ and n known, nb of states known (Abe & Warmuth, 1992)

Saint-Etienne

Proba-Poly-PAC (1/2)

Possible

- Proba-d_∞-PAC PFA (Angluin 88, lemma 14)
- Proba-KL-PAC Unigram with unknown vocabulary (Mc Allester & Shapire, 2000)
- Proba-KL-PAC PFA on Σ^n, Σ and n known, nb of states known (Abe & Warmuth, 1992)
- Proba-KL-PAC Acyclic PDFA on Σ^n, nb States known, Σ and n known (Ron & al., 1995)

Saint-Etienne

Proba-Poly-PAC (1/2)

Possible

- Proba-d_∞-PAC PFA (Angluin 88, lemma 14)
- Proba-KL-PAC Unigram with unknown vocabulary (Mc Allester & Shapire, 2000)
- Proba-KL-PAC PFA on Σ^n, Σ and n known, nb of states known (Abe & Warmuth, 1992)
- Proba-KL-PAC Acyclic PDFA on Σ^n, nb States known, Σ and n known (Ron & al., 1995)
- Proba-KL-PAC PDFA Cyclic Aut structure (Thollard & Clark, 2004)

Saint-Etienne

Proba-Poly-PAC (1/2)

Possible

- Proba-d_∞-PAC PFA (Angluin 88, lemma 14)
- Proba-KL-PAC Unigram with unknown vocabulary (Mc Allester & Shapire, 2000)
- Proba-KL-PAC PFA on Σ^n, Σ and n known, nb of states known (Abe & Warmuth, 1992)
- Proba-KL-PAC Acyclic PDFA on Σ^n, nb States known, Σ and n known (Ron & al., 1995)
- Proba-KL-PAC PDFA Cyclic Aut structure (Thollard & Clark, 2004)
- Proba-KL-PAC PDFA Cyclic Aut + informations (Clark & Thollard, 2004)

Saint-Etienne

Proba-Poly-PAC (2/2)

Impossible

- Proba AFN on Σ^n, Σ unknown
(Abe & Warmuth, 1992)
- *PDF of unknown class* on $\{0, 1\}^n$
(Kearns & al., 1994)
- Proba-KL-PAC Cyclic Aut, without aut information
(Clark & Thollard, 2002)

Saint-Etienne

Contents

- Probabilistic Finite State Machines
- What learning means
- Learnability results
- Algorithmic issues
  - Template technique
  - Instantiation of the template technique
  - Smoothing automata
- Experimental issues
- Conclusions

Saint-Etienne

Template algorithm

The common strategy follows two steps

- Building a maximum likelyhood estimate of the data

Saint-Etienne

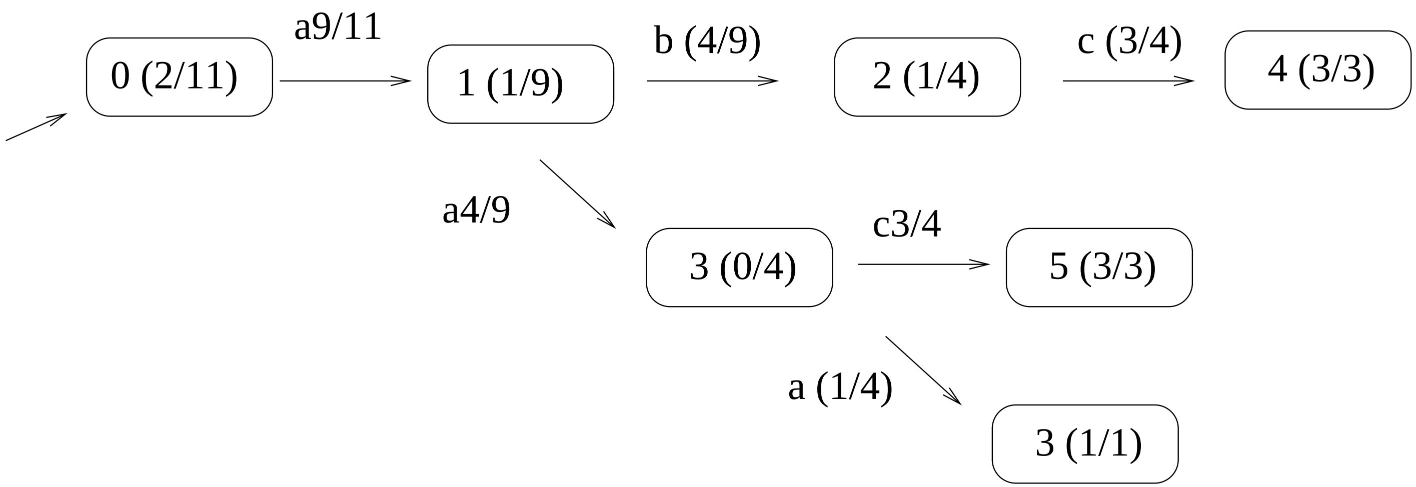
Template algorithm

The common strategy follows two steps

- Building a maximum likelyhood estimate of the data
- Generalizing using sate merging operations.

Saint-Etienne

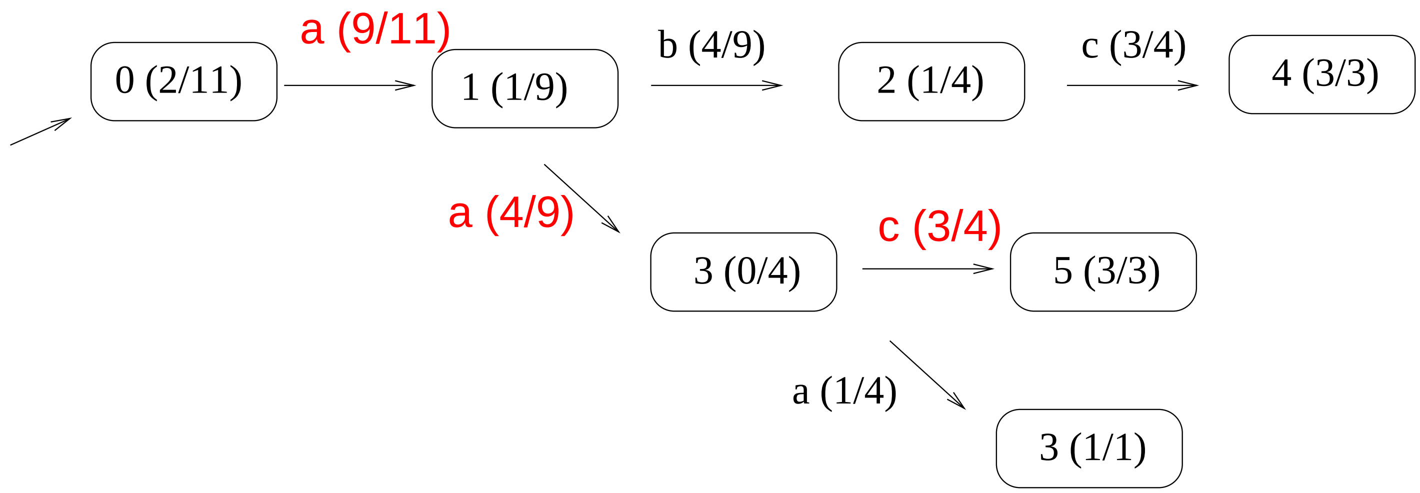
Learning by heart: the PPTA



PPTA of the learning set

EA = {λ,aac,aaa, aac, abc,aac,abc, abc, λ, a, ab }

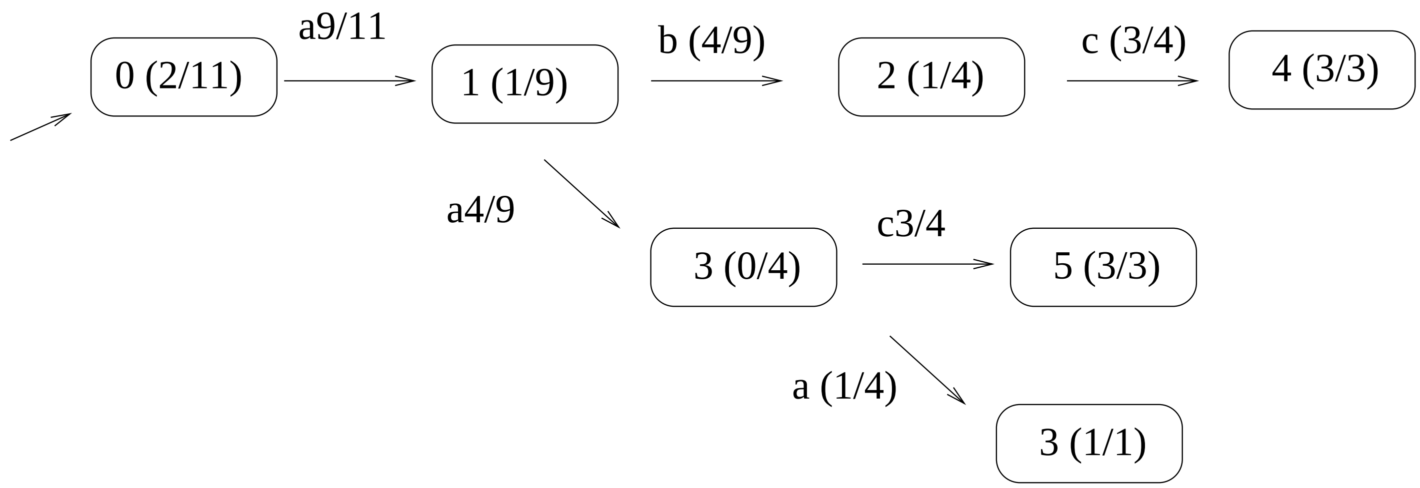
Learning by heart: the PPTA



PPTA of the learning set

EA = {λ,aac,aaa, aac, abc,aac,abc, abc, λ, a, ab }

Learning by heart: the PPTA



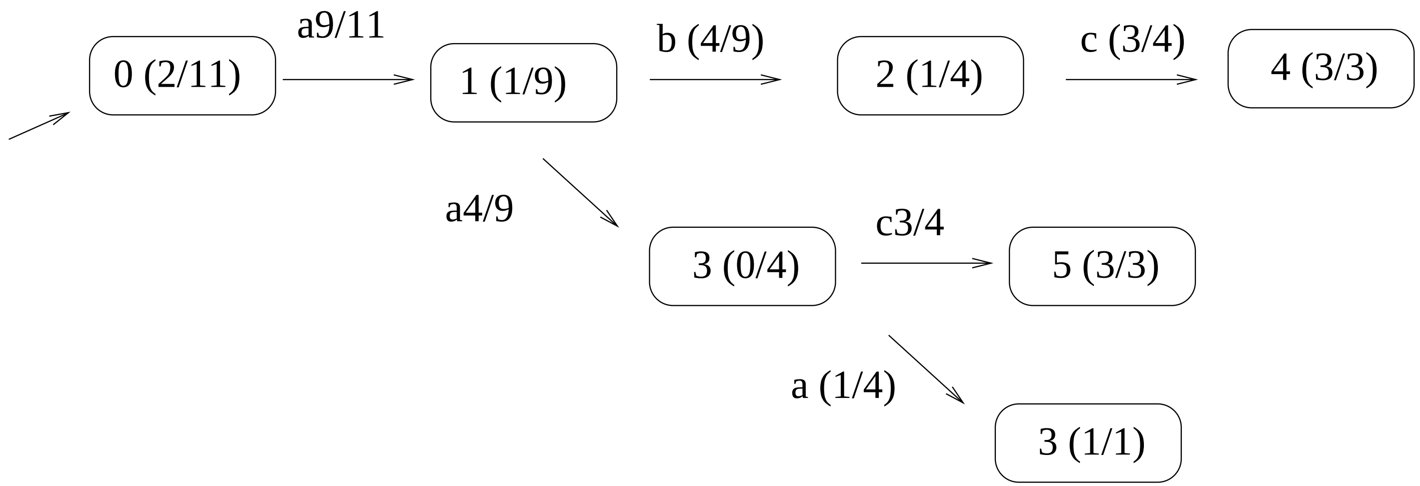
PPTA of the learning set

EA = {λ,aac,aaa, aac, abc,aac,abc, abc, λ, a, ab }

Note: String "aba" has null probability

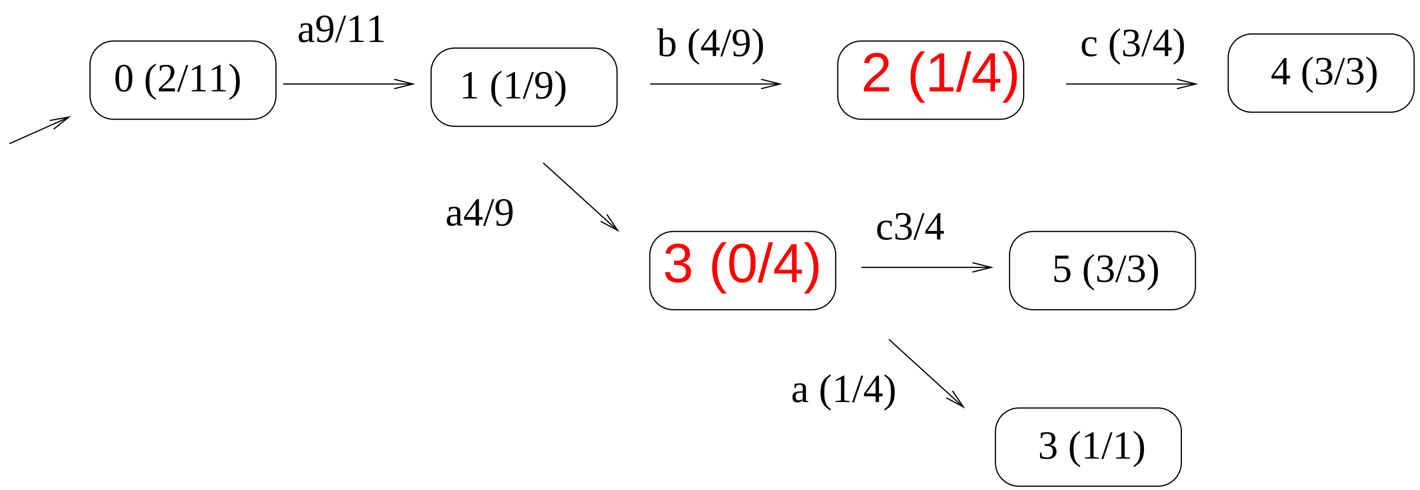
Generalization

Choosing two states



Generalization

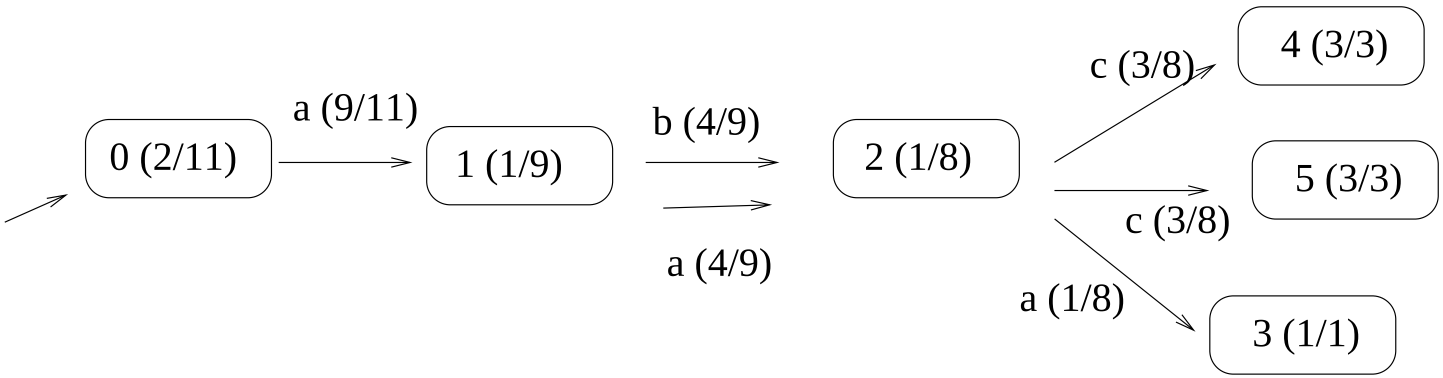
Choosing two states



Choosing state 2 and 3

Generalization

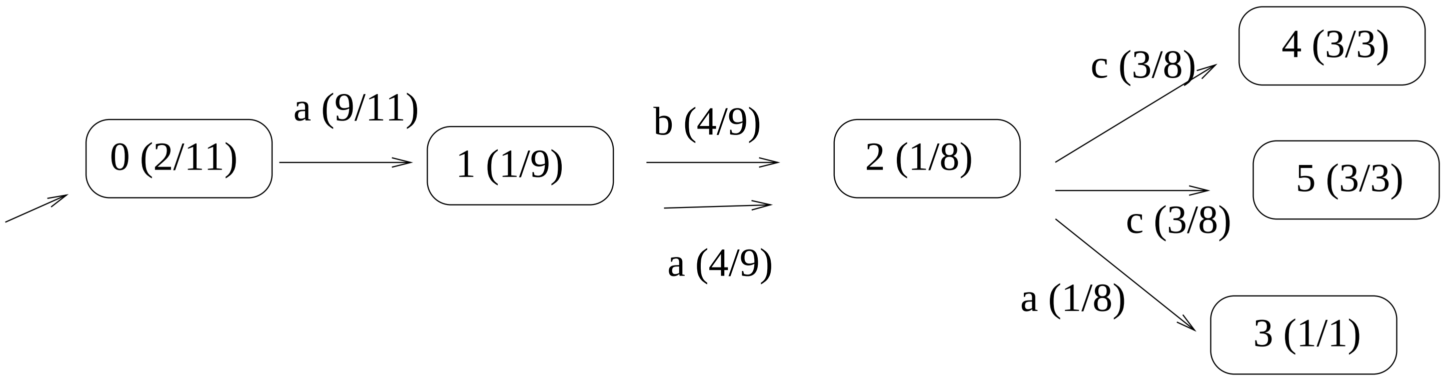
Merging states 2 and 3



Merging 2 and 3: can lead to non determinism

Generalization

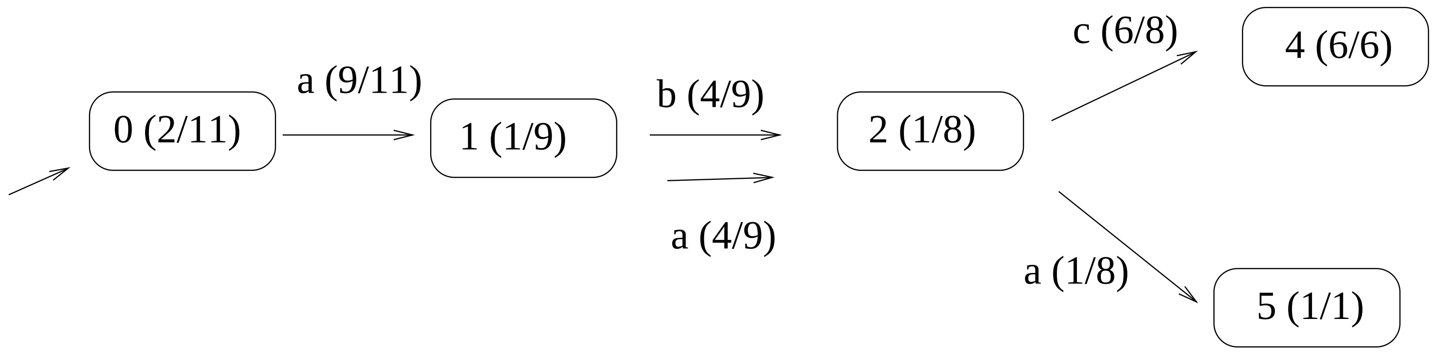
Merging states 4 and 5



Merging states 4 and 5

Generalization

After the "determinization"



Note: String "aba" has non null probability

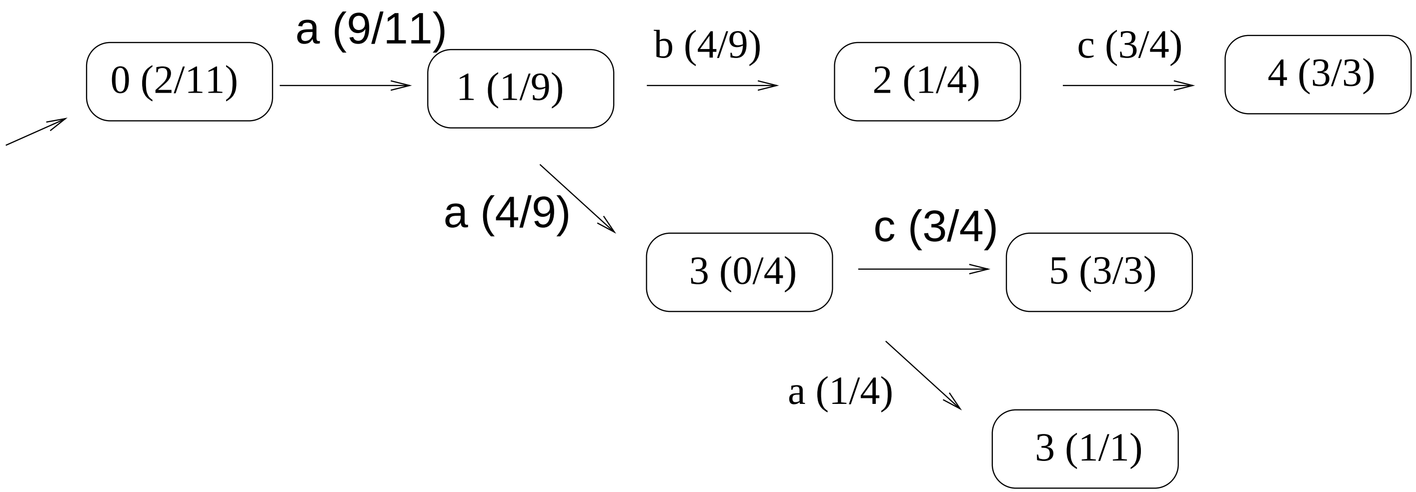
Generic Algorithm

Input: A multiset of string EA, a real α

Output: A PDFA

```
begin
    Building-PPTA (EA) ;
    A ← PPTA ;
    while (q_i,q_j) ← Choosing-Two-States (A) do
        if Compatible (i, j, α) then
            A ← Merge (A,q_i,q_j) ;
        end
    end
    Return A ;
end
```

Ordering merges



PPTA built on th multiset

EA = {λ, aac, aac, abc, aac, aac, abc, abc, λ, a, ab }

Merging ordering

Alergia (Carrasco & Oncina, 1994) :

HMM-infer (Stolcke, 1994): looking at **each** merge at **each** time $\rightarrow$ not tracktable on big data sets

LAPPTA (Ron & al., 1995): building of acyclic automata

EDSM (Lang, 1998): merging ordering based on the quantity of information

DDSM (Thollard,2001): ordering adapted from the EDSM algorithm.

Saint-Etienne

Compatibility tests

Alergia (Carrasco, 1994): Statistic test based on Hoeffding bounds

LAPPTA (Ron & al., 1995): Statistic test based on similarity measure

Youg-Lai & Tompa, (2000): Same as Alergia but emphasis on low frequency problem

MDI (Thollard & al., 2000): Tradeoff between size and distance to the data

M-Alergia (Kermorvant & Dupont, 2002): Statistical test based on multinomial test

Alergia (Habrard & al., 2003): Defines and deal with uniform noise.

Saint-Etienne

Other learning schemes

- Splitting/merging strategy (Brant, 1996)
- Incremental learning (Carrasco'99, Thollard & Clark, 2004, Callut & Dupont, 2004)

Saint-Etienne

The smoothing problem

The farm example

Let F be a farm with:

- 3 chickens
- 2 ducks

What is the probability of:

The smoothing problem

The farm example

Let F be a farm with:

- 3 chickens
- 2 ducks

What is the probability of:

Pr(chicken) = 3/5

The smoothing problem

The farm example

Let F be a farm with:

- 3 chickens
- 2 ducks

What is the probability of:

Pr(pig) = 0

Saint-Etienne

The smoothing problem

The farm example

Let F be a farm with:

- 3 chickens
- 2 ducks

What is the probability of:

Pr(lion) = 0

The smoothing problem

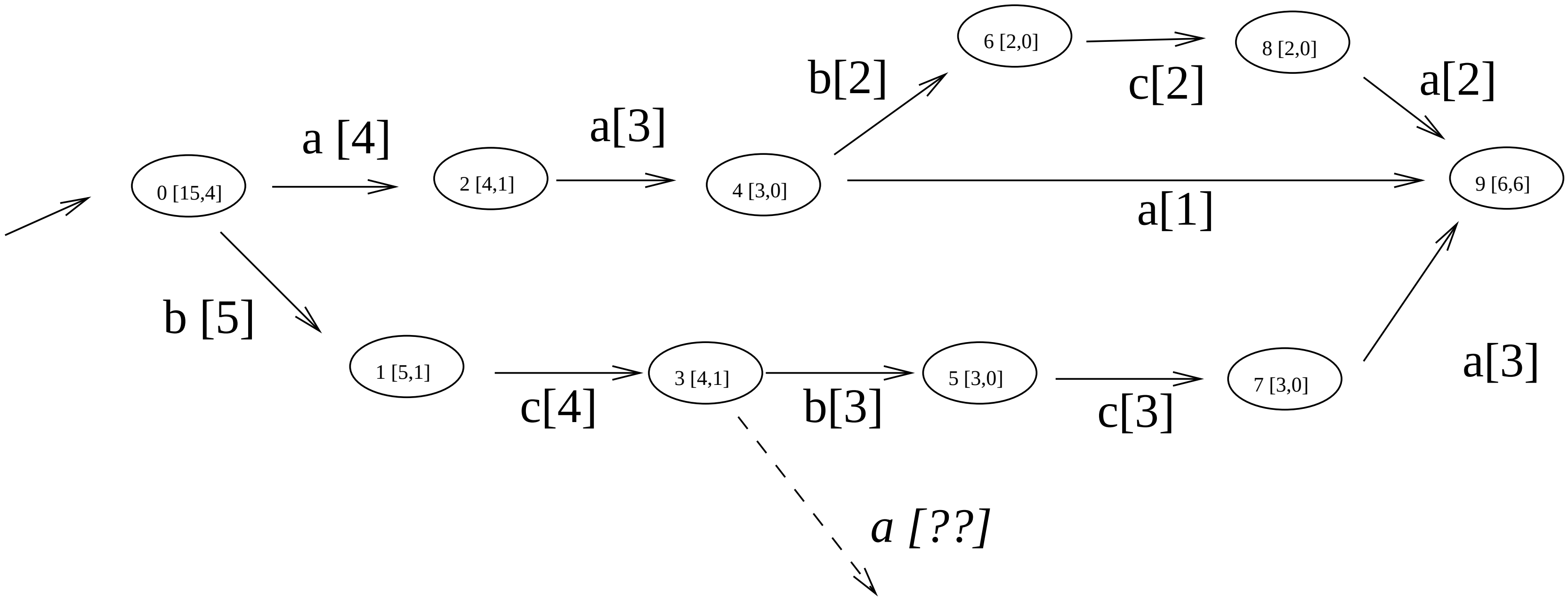
Considering the chain rule

n-grams the problem needs to be considered **only** with null probability n-grams.

Automata

- estimating the first null transition ?
- where to continue in the automaton ? in the same automaton ? on others ?

The smoothing problem



Smoothing automata

LAPPTA (Ron & al., 1995): Creation of a "small frequency state"

Alergia2 (Young-lai & Tompa, 2000): Emphasis on small frequency transitions

Error-correcting (Dupont & Amengual, 2000): Error correcting

Discounting (Thollard,2001): Back-off to a unigram

Discounting (Llorens & *al.*, 2002): Back-off to automata

Additive discounting (Thollard & Clark, 2004): Theoretical justification.

Discounting (Mc Allester & Shapire, 2000): Theoretical discounting for the unigram.

Saint-Etienne

Contents

Saint-Etienne

Experimental issues (MDI/DDSM)

- **Language Modeling:** much compact models (Thollard, 2001),

Experimental issues (MDI/DDSM)

- **Language Modeling:** much compact models (Thollard, 2001),
- **Speech to text:** faster at parsing time than bigrams (Exeperimentation at CNET),

Saint-Etienne

Experimental issues (MDI/DDSM)

- **Language Modeling:** much compact models (Thollard, 2001),
- **Speech to text:** faster at parsing time than bigrams (Exeperimentation at CNET),
- **Noun phrase chunking:** competitive results ($\sim$ 90 %) (Thollard & Clark, 2004),

Experimental issues (MDI/DDSM)

- **Language Modeling:** much compact models (Thollard, 2001),
- **Speech to text:** faster at parsing time than bigrams (Exeperimentation at CNET),
- **Noun phrase chunking:** competitive results ($\sim$ 90 %) (Thollard & Clark, 2004),
- **Body rule generation:** better and more compact than n-gram (Infante-Lopez, 2004).

Around the inference

Clustering (Dupont & Chase, 1998)

Interpolating automata (Thollard, 2001)

Bagging (Thollard & Clark, 2002)

Boosting (Thollard & al. 2002)

Typing automata (Kermorvant & de la Higuera, 2002)

Contents

Saint-Etienne

Conclusion

Probabilistic grammatical inference

- Framework in which theoretical results exist
- Good results in many domains (e.g NLP)
- Can deal with big data sets (e.g. Wall Street Journal)
- Provides very compact automata.

Open questions

Theoretical:

- Learning/smoothing n-gram models
- What is a good distance for the Proba-D-PAC framework ?

Practical:

- Learning non-deterministic models
- Improving the merging ordering
- Algorithmic: improving the algorithms