Stochastic grammatical inference with multinomial tests

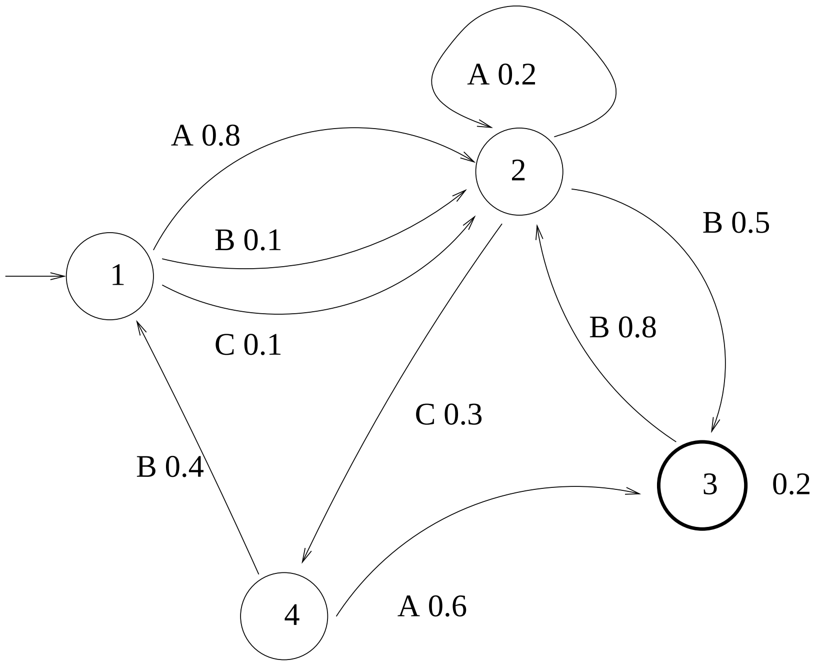
Christopher Kermorvant[1] and Pierre Dupont[2]

[1]EURISE, Université Jean Monnet, Saint-Etienne, France

[2]INGI, University of Louvain, Louvain-la-Neuve, Belgium

Framework (1)

Infer a stochastic regular language : stochastic automaton (SDFA)



Framework (2)

Use only positive examples :

- language modeling :
  WHAT IS THE CHEAPEST FLIGHT TO BOSTON.

  ROUND TRIP PLEASE.
- bio-informatics : ABC_tran :
  G Q I T S F L G H N G A G K ...
  G E C F G L L G V N G A G K ....

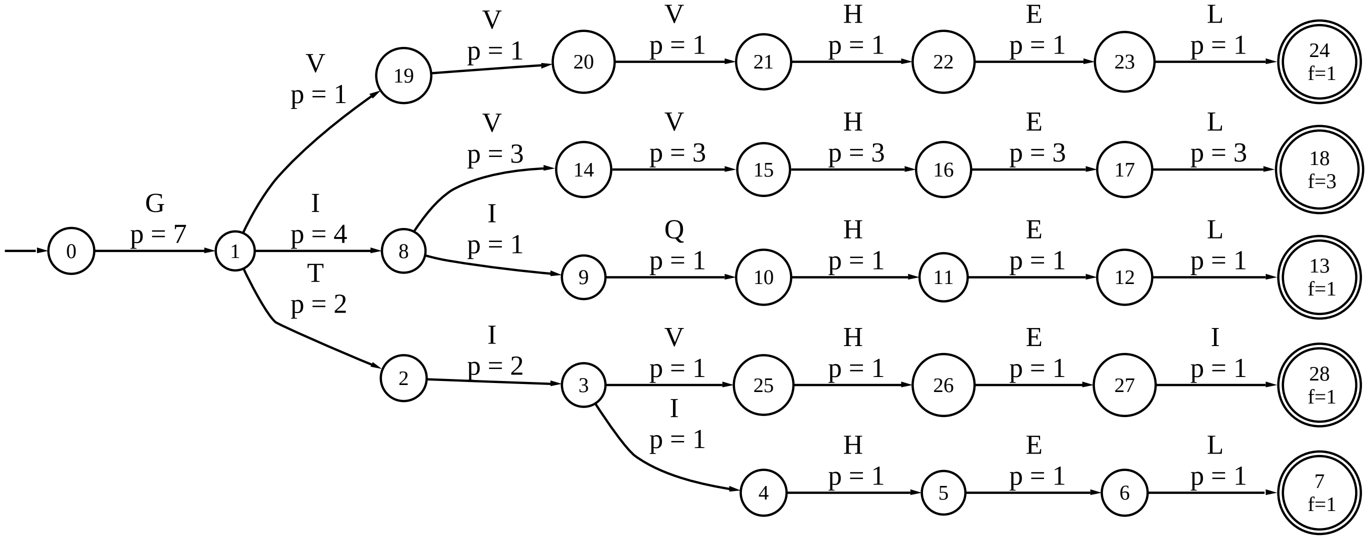
Consider practical cases, in particular small sample case.

Principles of merge based inference (1)

Consider all the examples : build the stochastic prefix tree acceptor (SPTA)

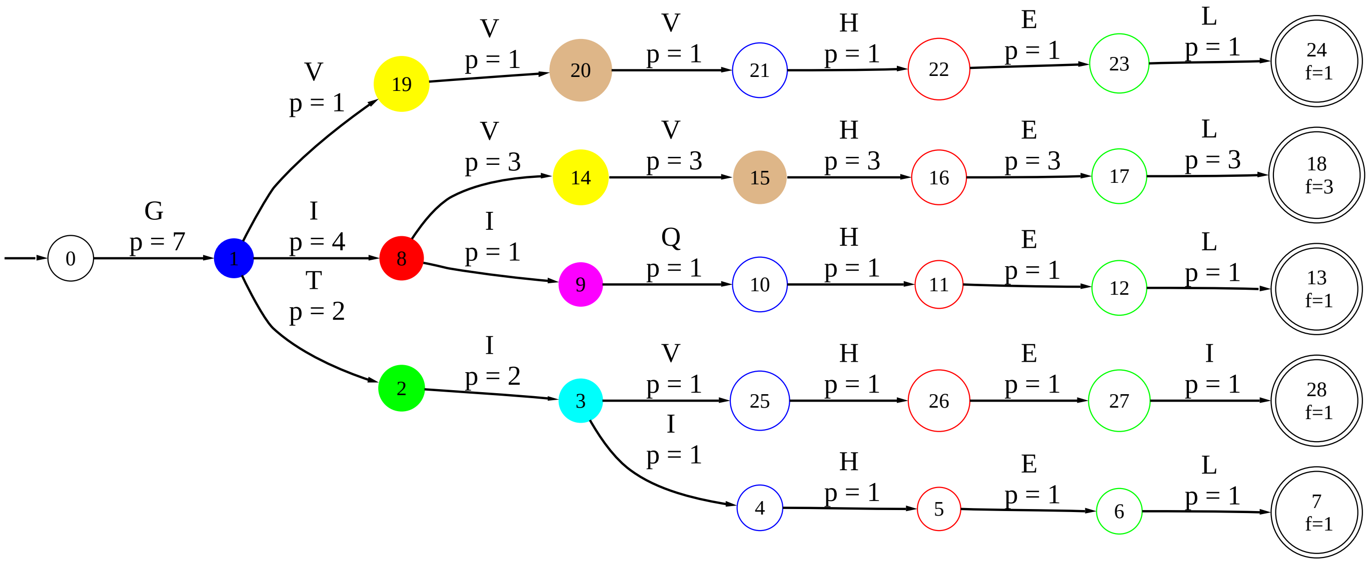
G T I I H E L
G I I Q H E L
G I V V H E L
G I V V H E L
G I V V H E L
G V V V H E L
G T I V H E I

⇒



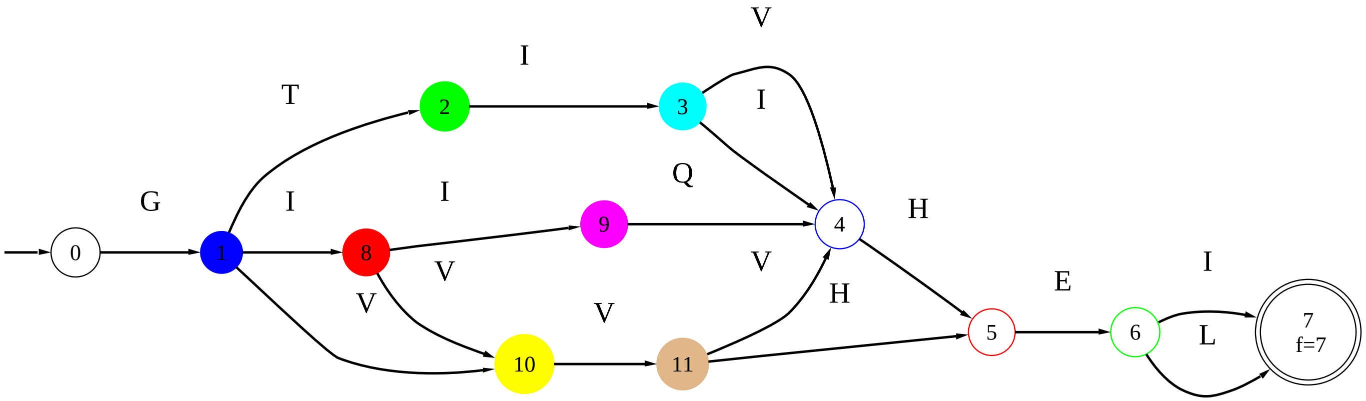
Principles of merge based inference (2)

Find *equivalent* states : corresponding to the same state in the target automaton



Principles of merge based inference (3)

Merge *equivalent* states



Equivalence of 2 states (1)

Alergia[Carrasco94], RLIPS[Carrasco99] :
Two states are equivalent iff

- for each letter, their transition probabilities are compatible
- their final probabilities are compatible
- their successors are compatible (recursively)

Equivalence of 2 states (2)

Two states q_1 and q_2 are compatible iff:

$$\forall a \in \Sigma, \left|\frac{C(q_1,a)}{C(q_1)} - \frac{C(q_2,a)}{C(q_2)}\right| < \sqrt{\frac{1}{2}\ln\frac{2}{\alpha}}\left(\frac{1}{\sqrt{C(q_1)}} + \frac{1}{\sqrt{C(q_2)}}\right) \quad (1)$$

$$\left|\frac{C_f(q_1)}{C(q_1)} - \frac{C_f(q_2)}{C(q_2)}\right| < \sqrt{\frac{1}{2}\ln\frac{2}{\alpha}}\left(\frac{1}{\sqrt{C(q_1)}} + \frac{1}{\sqrt{C(q_2)}}\right) \quad (2)$$

$$\forall a \in \Sigma, \delta_Q(q_1,a) \text{ and } \delta_Q(q_2,a) \quad \text{are compatible} \quad (3)$$

Equivalence of 2 states (3)

Limitations : probabilities compatibility test derived from Hoeffding bounds

1. no global test (one test for each transition)
2. asymptotic test : what about small sample case ?
3. impossible to compare two tests : which states are *more* compatible ?

Possible solution for 1 & 3 : MDI [Thollard et al. 2000]

Multinomial state model (1)

Each state is considered as a multinomial distribution

- parameters $\tau = \{\tau_1, \tau_2, \cdots, \tau_K\}$
- τ_i is the transition probability on the ith letter
- final probability included as a special symbol

Multinomial state model (2)

In the PPTA :

- each state is a realization of a multinomial distribution.
- equivalent states = states corresponding to the same underlying distribution.
- use statistical tests on multinomial distributions to test :

$$H_0 : q_1 \text{ and } q_2 \text{ are equivalent states.}$$

Multinomial state model (3)

Contingency table for two states under H_0 hypothesis (2 letters alphabet) :

	letter a	letter b	total
state q_1	$C(q_1, a)$ $H_{1a} = C(q_1)\hat{\tau}_a^q$	$C(q_1, b)$ $H_{1b} = C(q_1)\hat{\tau}_b^q$	$C(q_1)$
state q_2	$C(q_2, a)$ $H_{2a} = C(q_2)\hat{\tau}_a^q$	$C(q_2, b)$ $H_{2b} = C(q_2)\hat{\tau}_b^q$	$C(q_2)$
total	$C(a)$	$C(b)$	N

C are counts, $\hat{\tau}$ multinomial parameters estimated under H_0

Multinomial compatibility test (1)

- The Pearson statistic :

$$X^2 = \sum_{k=1}^{K} \left(\frac{(C(q_1,k) - H_{1k})^2}{H_{1k}} + \frac{(C(q_2,k) - H_{2k})^2}{H_{2k}} \right)$$

- $X^2 \sim \chi^2$ distribution with $K-1$ degrees of freedom
- H_0 rejected with confidence α if $X^2 > \chi^2(K-1,\alpha)$

note : correcting the paper

Multinomial compatibility test (2)

Small sample case : Fisher Test

- enumerate all the contingency tables with the same marginal counts as the tested table
- compute their probability by the hypergeometric distribution
- sum the probabilities of the tables at least as unfavorable to H_0
- directly compare this sum to the confidence threshold α to accept or reject H_0.

Multinomial compatibility test (3)

Algorithmic complexity :

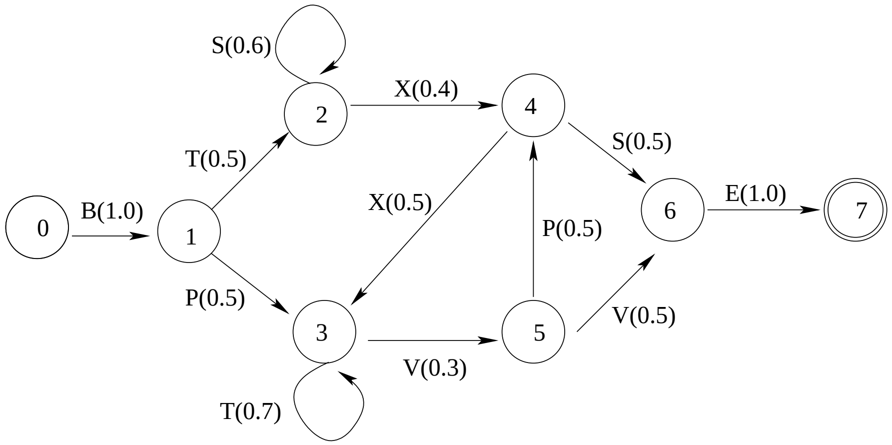
- χ^2 values are tabulated : same complexity as Alergia
- Fisher test exponential in alphabet size but :
  - use Fisher only with small counts
  - use approximations

Evidence driven strategies

- Adaptation of data dependent strategies (Blue-Fringe) to stochastic framework
- Do first merging operations supported by the most evidence
- Use p-value to compare states couples : the lower the p-value, the less compatible the states

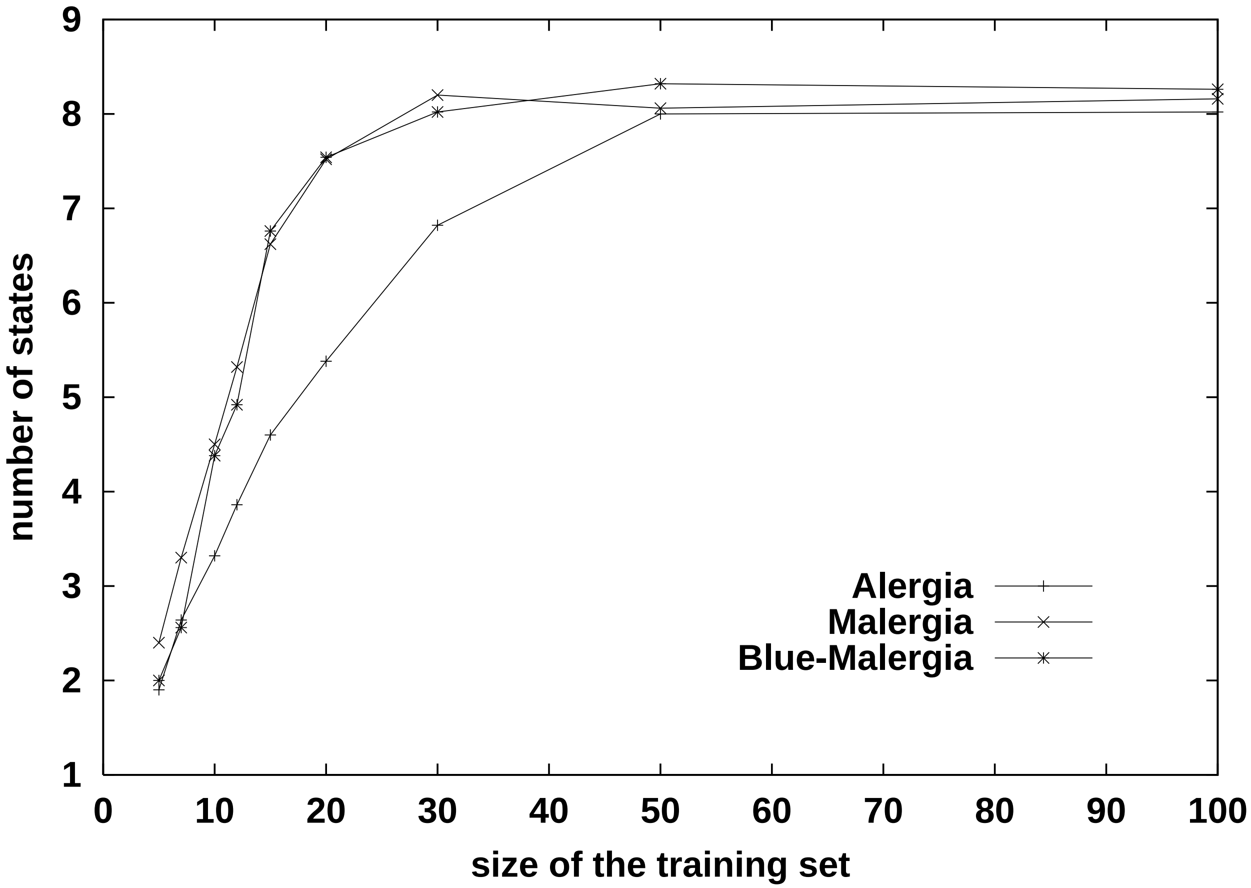
Experiments (1)

- Compare Alergia, Malergia and Blue-Malergia
- Test on the Reber task

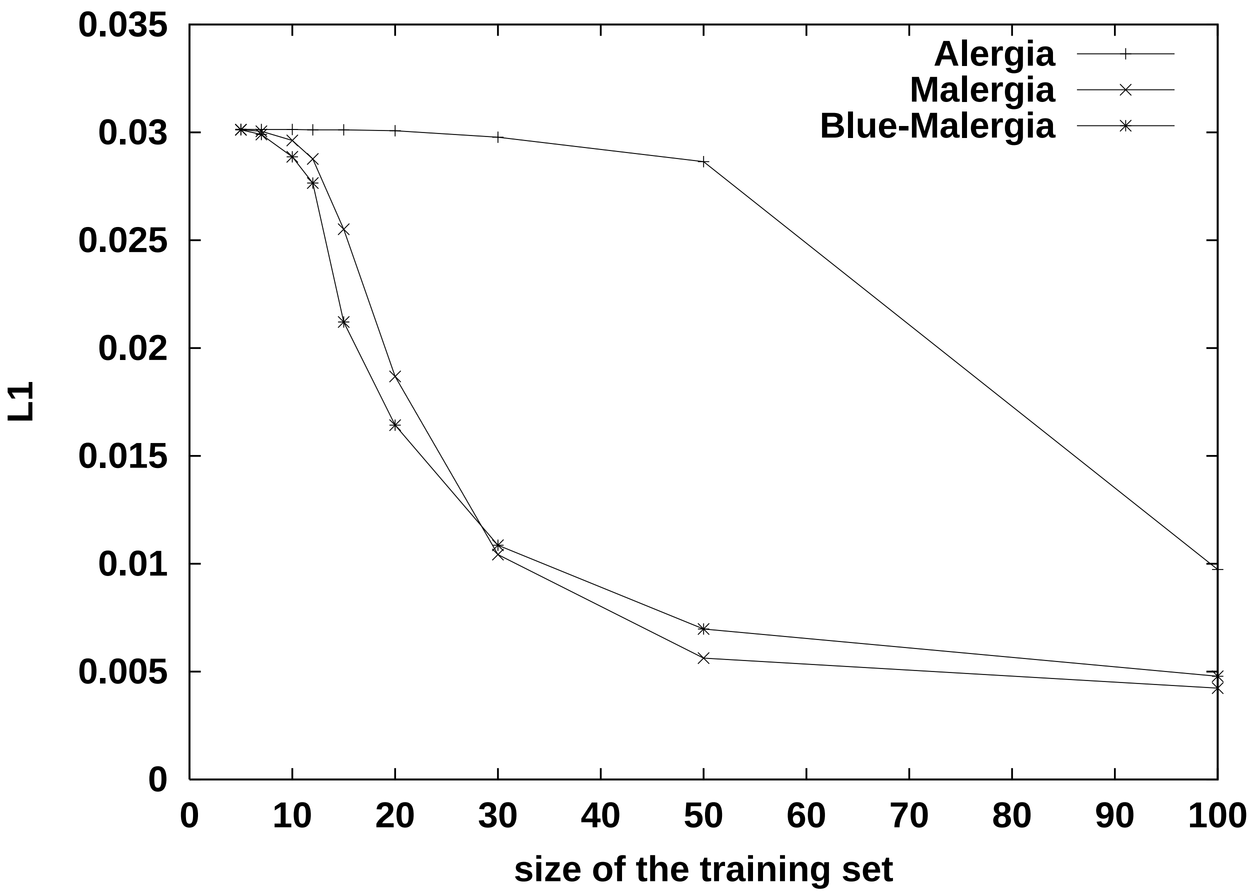


- Use $\bar{L1} = \frac{1}{|S|}\sum_{w\in S}|P_t(w) - P_i(w)|$ computed on 10 000 words ($\approx$ 100 000 symbols) averaged on 50 runs

Experiments (2)



Experiments (3)



Conclusion (1)

Malergia : statistical framework for inference

- model a state globally, with all its transitions
- deal with small sample cases
- allow data dependent strategies

Conclusion (2)

- Malergia : better convergence with less examples
  - quickest convergence to the right number of states
  - smallest distance $\bar{L1}$
- Advantage of data depend strategy not shown.
  - possible explanation : we focused on type I error; try to use type II error.

Perspective

- Better experimental validation
- Learnability in a statistical framework
- Applications