Learning Syntactic Structure from Examples and Applications

Enrique Vidal Ruiz

Instituto Tecnológico de Informática
Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Spain

evidal@iti.upv.es

October 2000

Index

1. ***Introduction***

2. Grammars, Automata and Formal Languages

3. Learning Paradigms and Fundamental Results

4. Learning Finite-State Automata

5. Learning Context-Free Grammars

6. Learning Finite-State Transducers

7. Applications

Sequences and Syntactic Structure

Examples of Syntacticly Structured sequences:

- Electrocardiogram
- Chain-coded image contour
- Banded chromosome grey-level sequence
- Acoustic-phonetic sequence
- String of characters
- Natural Language sentence
- Piece of music
- DNA sequence
- etc. etc. etc. . . .

▷ ***Syntactic Structure is properly represented by Grammars and Automata***

Learning Syntactic Structure and Grammatical Inference (GI)

WHAT IS ***GRAMMATICAL INFERENCE*** (GI)?

- Theory and Methods for Learning Grammars from Training Data.
- Well established discipline dating back to the 60's [Gold,1967]: "Language Identification in the Limit". (Information & Control).
- Traditional application field: Syntactic Pattern Recognition [Fu,1982].
- Many other (potential) applications including Genetic Sequence, Speech and Natural Language processing.

Literature on Grammatical Inference

Journals:

- Information & Computation (old Information & Control)
- Theoretical Computer Science
- Journal of the ACM
- Machine Learning
- IEEE Transactions on PAMI
- IEEE Transactions on SMC
- Int. Journal on Pattern Recognition and Artificial Intelligence

Conferences and Workshops:

- International Colloquium on Grammatical Inference (ICGI) (biannual)
- Computational Learning Theory (COLT) (annual)
- Algorithmic Learning Theory (ALT) (annual)
- International Conference on Pattern Recognition (ICPR) (biannual)
- Workshop on Syntactic & Structural Pattern Recognition (biannual)

Literature on Grammatical Inference

Books:

- R.González and M.Thomason: ”Syntactic Pattern Recognition. An Introduction”. Addison-Wesley 1978.
- K.S.Fu: ”Syntactic Pattern Recognition and Applications”. PrenticeHall 1982
- Osherson Stob Weinstein: ”Systems that Learn: an Introduction to Learning Theory for Cognitive and Computer Scientists”. MIT Press 1986.

Review Papers:

- K.S.Fu: ”Grammatical Inference: Introduction and Survey”. Parts 1 & 2 IEEE Trans. System Man and Cybernetics. Vol SMC-5 No.5 pp.95-11 409-423 1975.
- D.Angluin and C.H.Smith: ”Inductive Inference: Theory and Methods”. ComputingSurveys Vol.15 N**?**3 pp.46-62 1983.
- L.Miclet: ”Grammatical Inference”. In ”Syntactic and Structural Pattern Recognition and applications”. H.Bunke A.Sanfeliu (eds.) pp.237-290. World Scientific 1990.

Index

1. Introduction

2. ***Grammars, Automata and Formal Languages***

3. Learning Paradigms and Fundamental Results

4. Learning Finite-State Automata

5. Learning Context-Free Grammars

6. Learning Finite-State Transducers

7. Applications

Grammars

- ***Free Monoid*** Σ^****:*** Given a finite set Σ, Σ^+ is the set of all finite-length strings of elements of Σ, and $\Sigma^* = \Sigma^+ \cap \{\lambda\}$ (the empty string).

- ***Grammar:*** $G = (N, \Sigma, R, S)$
  - N: finite set of *Non-Terminals*
  - Σ: finite set of *Terminals or Primitives*
  - $S \in N$: Starting Non-Terminal or *"Axiom"*
  - $R \subset (N \cup \Sigma)^* N (N \cup \Sigma)^* \times (N \cup \Sigma)^*$: set of *Rules*.

  A rule is written as:

  $$\boldsymbol{a \longrightarrow b}, \qquad a \in (N \cup \Sigma)^* N (N \cup \Sigma)^*, \;\; b \in (N \cup \Sigma)^*$$

Grammars and Languages

- ***Elementary Derivation:*** $\underset{G}{\Longrightarrow}$***:***

  $\mu\,\alpha\,\delta \underset{G}{\Longrightarrow} \mu\,\beta\,\delta$ iff $\exists(\alpha \rightarrow \beta) \in R,\ \ \mu, \delta \in (N \cup \Sigma)^*$

- ***Derivation*** $\overset{*}{\underset{G}{\Longrightarrow}}$***:***

  Is a *finite sequence of elementary derivations*. A derivation d can be written as the corresponding sequence of rules of G.

  The *set of derivations* of $x \in \Sigma^*$ (such that $S \overset{*}{\underset{G}{\Longrightarrow}} x$) is written as $D_G(x)$.

  A grammar G is *ambiguous* if $\exists x \in \Sigma^*$ such that $|D_G(x)| > 1$

- ***Language generated by a grammar*** $G,\ \mathcal{L}(G)$***:***

  $$\mathcal{L}(G) \ = \ \{\, x \in \Sigma^* \ \mid \ S \overset{*}{\underset{G}{\Longrightarrow}} x \,\}$$

Types of Grammars and Languages

CHOMSKY HIERARCHY FOR RECURSIVE LANGUAGES

0: Unrestricted

1: Context Sensitive

$\alpha B \gamma \rightarrow \alpha \beta \gamma \ , \qquad \alpha, \gamma \in (N \cup \Sigma)^*, \ \ B \in N, \ \ \beta \in (N \cup \Sigma)^+$

2: ***Context Free***

$B \rightarrow \beta \ , \qquad B \in N, \ \ \beta \in (N \cup \Sigma)^*$

3: ***Regular or Finite-State***

$A \rightarrow aB$ or $A \rightarrow a \ , \quad A, B \in N, \ \ a \in \Sigma \cup \{\lambda\}^*$

Regular Grammars and Finite-State Automata

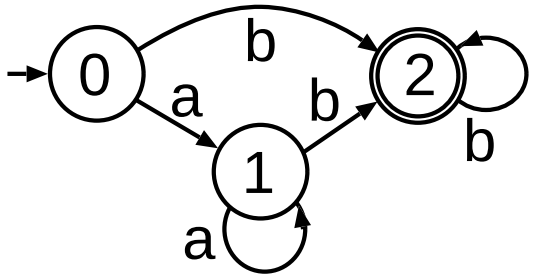
- ***Regular Grammars:*** $G = (N, \Sigma, R, S)$,
  $(A \rightarrow aB) \in R \vee (A \rightarrow a) \in R, \;\; A, B \in N, \;\; a \in \Sigma$

- ***Finite-State Automata:*** $\mathcal{A} = (Q, \Sigma, \delta, q_0, F), \quad q_0 \in Q, \;\; \delta : \; Q \times \Sigma \; \rightarrow \; 2^Q$

- ***Equivalence:*** For every Regular Grammar there exists a Finite-State Automaton that recognizes the same language. (*The converse is not always true for* **stochastic** *languages !*).

Example:

$G = (N, \Sigma, R, S);$
$\Sigma = \{a, b\}; \; N = \{S, A_1, A_2\};$
$R = \{ \; S \; \rightarrow aA_1 \mid bA_2 \mid b,$
$\quad A_1 \rightarrow aA_1 \mid bA_2 \mid b,$
$\quad A_2 \rightarrow bA_2 \mid b\}$



$\mathcal{A} = \{Q, \Sigma, \delta, q_0, F\};$
$Q = \{0, 1, 2\},$
$\Sigma = \{a, b\},$
$q_0 = 0, \; F = \{2\}$

Stochastic Grammars and Languages

- A ***Stochastic Grammar*** G' is a grammar G with probabilities associated to its rules:

$$G' = (G, p), \quad G = (N, \Sigma, R, S), \quad p : R \rightarrow [0, 1]$$

- A *context-free* (or *regular*) Grammar is ***proper*** if:

$$\forall A \in N \quad \sum_{A \rightarrow \beta \in R} p(A \rightarrow \beta) = 1$$

- ***Probability of a string*** x generated by G':

$$\forall x \in \Sigma^* \quad p(x|G') = \sum_{d \in D_G(x)} p(d), \qquad p(d) = \prod_{(A \rightarrow \beta) \in d} p(A \rightarrow \beta)$$

- A stochastic grammar G' is ***consistent*** if:

$$\sum_{x \in \Sigma^*} p(x|G') = 1$$

Estimating the Probabilities of the Rules

- ***Non-ambiguous Context-Free (or Regular) grammar G':***
  Maximum Likelihood estimation from frequencies of use of the rules in the parsing of a training set that is assumed to have been generated by G'.

  Estimates approach the true probabilities as the size of the training set $\rightarrow \infty$. [Mariansky et al., 77] [Chandhuriet al., 86].

- ***Ambiguous Context-Free (or Regular) grammars:***
  Locally optimal estimation using Backward-Forward or Inside-Outside re-estimation.

Index

1. Introduction

2. Grammars, Automata and Formal Languages

3. ***Learning Paradigms and Fundamental Results***

4. Learning Finite-State Automata

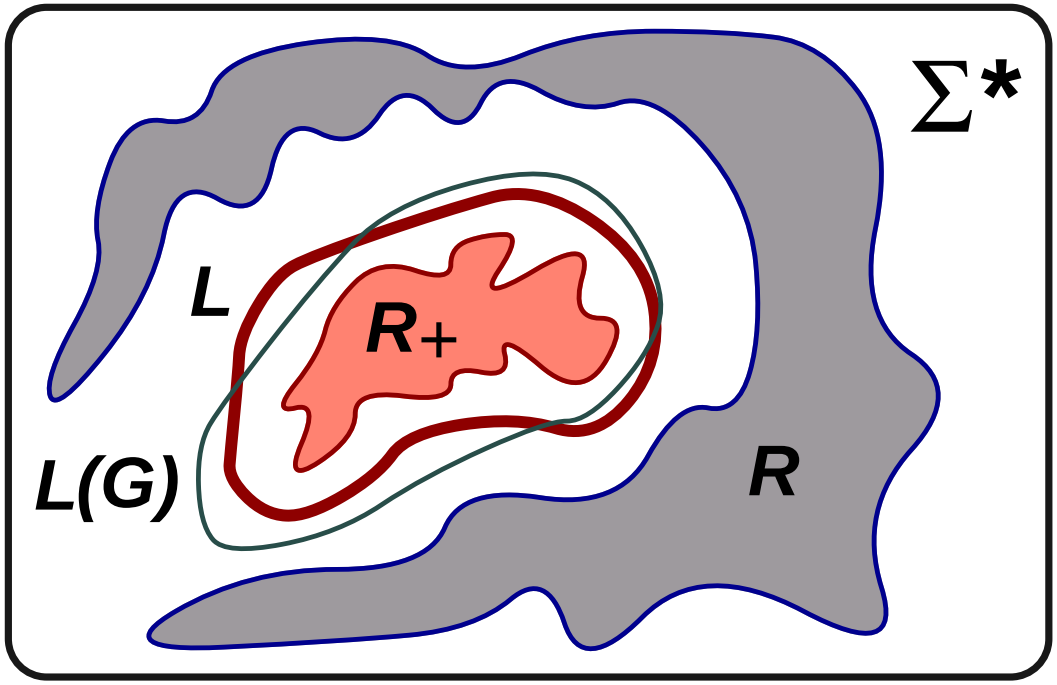
5. Learning Context-Free Grammars

6. Learning Finite-State Transducers

7. Applications

Language Identification in the Limit

- Given a *positive training set of an unknown language* L: $S_+ \subset L$
- [And possibly a *negative training set*: $S_- \subset \Sigma^* - L$]
- find a grammar G such that: $S_+ \subset \mathcal{L}(G)$ [and $S_- \cap \mathcal{L}(G) = \emptyset$]
- and, in the limit, for sufficiently large, $|S_+ \cup S_-|$: $\mathcal{L}(G) = L$



Computability Results

+ Any class of *recursively enumerable languages can be identified in the limit from* ***complete presentation*** (both positive and negative samples) [Gold,67].

However: Negative samples are not always available and also their use can lead to probably intractable problems [Angluin,78] [Gold,78].

– *No superfinite class of languages can be identified in the limit from only* ***positive presentation*** [Gold,67].

Therefore: Not even the class of Regular Languages is identifiable in the limit from only positive samples.

Possible escape: *use only positive samples* and:

- infer non-superfinite classes of languages: (***"characterizable" methods***)
  ... or
- supply adequate a-priori knowledge: (***"heuristic" methods***)

Using Positive Examples only

- Undecidability results strictly apply to superfinite classes. But many non-superfinite classes may exist that are identifiable with just positive information. E.g. undecidability does not apply to classes that do not contain all finite languages.

- The danger is *over-generalization*: if in a given inference step a language is produced that is larger than the unknown target language, this is irreversible since no positive sample could ever supply information to detect this error.

- By allowing only ***"conservative" generalizations*** over-generalization can be avoided for non-trivial classes of languages.

  An explicit characterization of these classes is given in [Angluin,80]

Design of GI Algorithms using Positive Data only

- ***Heuristic***: Rely on a *priori knowledge* on each application domain to avoid improper generalizations for this application.

  Examples:

  - K-Tails [Biermann & Feldman, 72]
  - Tail Clustering [Miclet,80]
  - Error Correcting Grammatical Inference (ECGI) [Rulot & Vidal, 87]
  - etc. . . .

- ***Characterizable***: Consider *specific classes of languages* which are identifiable in the limit.

  Examples:

  - K-Reversible Languages [Angluin,82]
  - Terminal Deterministic Languages [Radhakrishnan & Nagaraja, 88]
  - K-Testable Languages [García & Vidal, 90]
  - etc. . . .

  The (practical) problem consists of how to establish a relation between each considered application and a convenient class of (inferable) languages.

- ***A methodological tradeoff (MGGI):*** *techniques based on symbol relabeling functions (morphismes) [García Vidal & Casacuberta, 87*]

Other Learning Paradigms with more Optimistic Computational Results

- ***Approximate Identification*** in the limit is possible for ”dense” classes of Languages using only positive presentation [Wharton,74].

  An example is the identification of ***Stochastic Regular Languages*** [Carrasco & Oncina, 99].

- ***Other Helpful Learning Environments*** [Parekh & Hanovar, 2000]
  - *Probably Approximately Correct Identification (PAC)*
  - Learning from *Example Based Queries*
  - Using *Teaching and Characteristic Sets*
  - Learning from *Simple Examples*

Index

1. Introduction

2. Grammars, Automata and Formal Languages

3. Learning Paradigms and Fundamental Results

4. Learning Finite-State Automata

5. Learning Context-Free Grammars

6. Learning Finite-State Transducers

7. Applications

References to specific work

- [J.C.Amengual, J.M.Benedí, F.Casacuberta, A.Castaño, A.Castellanos, V.Jimenez, D.Llorens, A.Marzal, M.Pastor, F.Prat, E.Vidal, J.M.Vilar]: “The EUTRANS-I Speech Translation System”. Machine Translation, In press, 2000.
- [Angluin, 80]: D.Angluin: ”Inductive Inference of Formal Languages from Positive Data” Information and Control N.45, pp.117-135 1980.
- [Angluin, 78]: D.Angluin: ”On the Complexity of Minimun Inference of Regular Sets” Inf. and Control 39 pp. 337-350 1978.
- [Angluin, 82]: D.Angluin: ”Inference of Reversible Languages” J.ACM, 293 pp.741-765.
- [Angluin, 83]: D.Angluin and C.H.Smith: ”Inductive Inference: Theory and Methods”. Computing Surveys Vol.15 No.3 pp.46-62 1983.
- [Baker, 79]: J.K.Baker: ”Trainable Grammars for Speeach Recognition” In Speach Communication Papers for the 97th Meeting of the Accoustical Society of America (Klatt D. H. and Wolf J. J. eds). pp. 547-550.
- [Berstel, 79]: J.Berstel. ”Transductions and Contxt-Free Langages”. B. G. Teubner Stuggrt 1979.
- [Berwick, 87]: R.C.Berwick and S.Pilato: ”Learning Syntax by Automata Induction”. Machine Learning No.2 pp. 9-38 1987.
- [Biermann and Feldman, 72]: A.W.Biermann and J.A.Feldmann: ”On the Synthesis of Finite-State Machines from Samples of their behavior”.IEEE Trans. Compt. C-21 pp.592-597.
- [Brown, 90]: P.F.Brown et al. : ”A Stocastical Approach to Machine Translation” Computacional Linguistics Volume 16 Number 2 1990.
- [Casacuberta, 90]: F.Casacuberta E.Vidal B.Mas H. Rulot: ”Learning the Structure of HMMÕs Trrough Grammatical Inference Techniques”. IEEE-ICASSP proc. pp.717-720 April 1990.
- [Casacuberta, 90]: F.Casacuberta: ”Some Relations Among Stochastic Finite State Networks Used in Automatic Speech Recognition”. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol.12 No. 7.July 1990.
- [A.Castellanos, E.Vidal, A.Varó, J.Oncina]: “Language Understanding and Subsequential Transducer Learning”. Computer Speech and Language, No.12, pp.193-228. 1998.
- [Chandhuri et al., 86]: R.Chanhuri A.N.V.Rhao. ”Aproximating Grammar Probabilities]: Solution of a Conjecture”. Journal of ACM Vol-33N.4 pp.702-705 1986.
- P.P.Cruz, E.Vidal: “A Study of Grammatical Inference Algorithms in Automatic Music Composition”. En VII National Symposium on Pattern Recognition and Image Analysis. Pub. Asociación Española de Reconocimiento de Formas y Análisis de Imágenes (AERFAI). A. Sanfeliu, J. Villanueva and J. Vitriá (Eds). Barcelona. 1997.

- P.P.Cruz, E.Vidal: “Learning Regular Grammars to Model Musical Style. Proc. of 4th. International Colloquium on Grammatical Inference (ICGI-98). Springer-Verlag (Lecture Notes in Artificial Intelligence), pp.211-222. 1998.
- [Diaz, 93]: J. E. Diaz A. J. Rubio A. M. Peinado E. Segarra N. Prieto and F. Casacuberta. ”Development of task oriented Spanish Speech Corpora” Eurospeech 93 Proc. 1993.
- [Feldman, 90]: J.A.Feldman G.Lakoff A.Stolke and S.Hollbach Weber: ”Miniature Language Acquisition: A touchstone for cognitive science” International Computer Science Institute. TR-90-009. 1990.
- [Fu, 75]: K.S.Fu: ”Grammatical Inference: Introduction and Survey”. Parts 1 & 2 IEEE Trans. System Man and Cybernetics. Vol SMC-5 No.5 pp.95-11 409-423 1975.
- [Garcia Vidal and Casacuberta, 87]: P.Garcia E.Vidal and F.Casacuberta: ”Local Languages the Successor Method and a step towards a General methodology for the Inference of Regular Grammars”. IEEE Trans. on PAMIVol.9 No.6 pp.841-845. 1987.
- [Garcia, 90]: P.Garcia and E.Vidal: ”Inference of K-Testable Languages In the Strict Sense and Application to Syntactic Pattern Recognition”. IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. PAMI-12 N. 9 pp.920-925 1990.
- [Gold, 67]: E.M.Gold: ”Language Identification in the Limit”. Information and Control Vol.10 pp.447-474 1967.
- [Gold, 78]: E.M.Gold: ”Complexity of Automaton Identification from Given Data”. Inf. and Control37 pp.302-320 1978.
- [Guerrero, 92]: A.Guerrero E.Segarra and P.Garcia.: ”Utilizacion de un Modelo de Error basado en la Extension Semicontinua del Algoritmo k-EE en tareas de Reconocimiento del Habla”. V Simposium de Reconocimiento de Formas y Analisis de Imagenes pp.196-203 1992.
- [Jelinek, 91]: F.Jelinek: ”Up from Trigrams: The Stuggle for Improved Language Models” Eurospeech 1991.
- [Lari, 90]: K.Lari and S.J.Young: ”The Estimation of Stochastic context-free grammars using the Inside-Outside Algorithm” Cambrige University (U.K). Computer Speech and Language 4pp.35-361990.
- [Lari, 91]: K.Lari and S.J.Young: ”Aplications of Stochastic Context-Free Grammars Using the Inside-Outside Algorithm”. Computer Speech and Languages No.5 pp. 237-257 1991.
- [Mäkinen, 92]: E.Mäkinen: ”On the Structural Grammatical Inference Problem for some Classes of Context-Free Grammars”. Information Processing Letters 41. pp.1-5 1992.
- [Maryanski, 77]: F.J.Maryanski T.L.Booth. ”Inference of Finite-State Probabilistic Grammars”. IEEE Transactions on Computation Vol.C-26pp.531-5361977.
- [Medvedev, 64]: T.Yu.Medvedev: ”On the Class of Events Representable in a Finite Automaton” (translated from Russian) in Sequential Machines-Selected Papers ed. E. F. Moor Addison-Wesley pp.227-315. 1964.
- [Miclet, 80]: L.Miclet: ”Regular Inference with a Tail-Clustering Method”. IEEE Trans. Syst. Man Cyber. SMC-10 pp.737-743. 1980.

- [Miclet, 90]: L.Miclet: ”Grammatical Inference”. In ”Syntactic and Structural Pattern Recognition and applications”. H.Bunke A.Sanfeliu (eds.) pp.237-290. World Scientific 1990.
- [Muggleton, 84]: S.Muggleton: ”Induction of Regular Languages from Positive Examples ”. Tech. Rep. Turing Institute Research Memoranda Glasgow.
- [Oncina , 92]: J.Oncina P.Garcia and E.Vidal: ”Transducer Learning in Pattern Recognition” ICPR. Proc. 1992.
- [Oncina, 91]: J.Oncina: ”Aprendizaje de Lenguajes Regulares y Funciones Subsecuenciales” Universidad Politecnica de Valencia. 1991.
- [Oncina, 92]: J.Oncina and P.Garcia: ”Inferring Regular Languages in Polynomial Updated Time”. In Pattern Recognition and Image Analysis. N. Perrez de la Blanca A.Sanfeliu and E. Vidal (eds) Series in Machine Perception and Artificial Intelligence Vol-1. pp.49-61. World Scientific Plub 1992.
- [Oncina, 93]: J.Oncina P.Garcia and E.Vidal: ”Learning Subsequential Transducers for Pattern Recognition Interpretation Tasks”. IEEE Trans. on PAMI May 1993.
- [Pao, 78]: T.W.Pao and J.W.Carr:: ”A Solution of the Syntactical Induction-Inference Problem for Regular Languages”. Comput. Lan. 3 pp.53-64. 1978.
- [Pereira, 92]: F. Pereira and Y. Schabes: ”Inside-Outside Reestimation from Partially Bracketed Corpora”. 30 Annual Meeting of the Assocition for Computer Linguistic Proc. pp. 128-135 1992.
- [Pieraccini, 93]: R.Pieraccini and E.Levin: ”A Learning Approach to Natural Language Understanding” NATO-ASI on ASR 1993.
- [Price, 90]: P.J.Price: ”Evaluation of Spoken Language Systems: the ATIS Domain”. Proc. of 3rd. DARPA Workshop on Speech and Natural Language pp. 91-95 Hidden Valley (PA) June 1990.
- [Prieto, 92]: N.Prieto and E.Vidal: ”Learning Languages Model Through the ECGI method”. Speech Comunication 11 pp. 299-309 1992.
- [Radhakrishnan and Nagaraja, 88]: V.Radhakrishnan and G.Nagaraja: Inference of Even Linear Grammars and its Applications to picture Description Languages”. Pattern Recognition Vol.21 No.1 55-62.
- [Radhakrishnan , 87]: Radhakrishnan and G.Nagaraja: ”Inference of Regular Grammars via Skeletons”. Systems Man and Cybernetics 17 (6)pp. 11-21 1987.
- [Rulot and Vidal, 87]: H.Rulot and E.Vidal: ”Modelling (sub) string-lenght-based constraints through a Grammatical Inference Method” Devijver and Kittler eds. (Springer Berlin).
- [Rulot, 89]: H.Rulot N. Prieto E.Vidal: ”Learning Accurate Finite-State Structural Models of Words through the ECGI algorithm”. Proceedings ICASSP-89 Vol.1pp. 643-646 1989.

- [Rulot, 92]: H.Rulot : ”ECGI:Un Algoritmo de Inferencia Gramatical Mediante correccion de Errores”. Universitat de València. 1992.
- [Sakakibara, 92]: Y.Sakakibara : ”Efficient Learning of Context-Free Grammars from Positive Structural Examples”. Information and Computation 97. pp.23-60 1992.
- [Segarra, 93]: E.Segarra: ”Una Aproximacion Inductiva a la Comprension del Discurso Continuo”. Universidad Politecnica de Valencia. 1993.
- [Stolke, 93]: A.Stolke and S.Omohundro: ”Hidden Markov Model Induction by Bayesian Model Merging”. To appear in: C.L. Giles S.J. Hanson and J.D. Cowan eds. Advances in Neural Information Processing Systems 5 San Mateo CA Morgan Kaufman 1993.
- [Takada, 88]: Y. Takada: ”Grammatical Inference for Even Linear Languages Based on Control Sets”. Machine Learning pp. 375-377 1988.
- [Vidal, 93]: E.Vidal H.Rulot J.M.Valiente G.Andreu.: ”Application of the Error-Correcting Grammatical Inference Algorithm (ECGI) to Planar Shape Recognition”. Int. Col.loquium on Grammatical Inference]: Theory Applications and Alternatives. Proc. Colchester (U.K) 1993.
- [Vidal, 92]: E.Vidal H.Rulot J.M.Valiente G.Andreu : ”Font-Independent Mixed-Size Digit Recognition Through the Error-Correcting Grammatical Inference Algorithm (ECGI)”. ICPR-92 Proc. pp.334-337. 1992.
- E.Vidal: ”Language Learning, Understanding and Translation“. En “Progress and Prospects of Speech Research and Technology”, pp.131-140. H.Niemann, R.de Mori, G.Hanrieder (Eds.). Infix, 1994.
- J.M.Vilar, E.Vidal, J.C.Amengual: “Learning Extended Finite-State Models for Language Translation”. Proc. of Extended Finite State Models Workshop (of ECAI’96), pp.92-96. Budapest, Agosto 1996.
- E.Vidal and M.J.Castro: “Classification of Banded Chromosomes using Error-Correcting Grammatical Inference (ECGI) and Multilayer Perceptron (MLP)”. In VII National Symposium on Pattern Recognition and Image Analysis. Pub. Asociación Española de Reconocimiento de Formas y Análisis de Imágenes (AERFAI). A. Sanfeliu, J. Villanueva and J. Vitriá (Eds). Pages 31-36. Barcelona. 1997.
- E.Vidal: “Finite-State Speech-to-Speech Translation”. Int. Conf. on Acoustics Speech and Signal Processing (ICASSP-97), proc., Vol.1, pp.111-114. Munich, 1997.
- [Wharton, 74]: R.M.Wharton: ”Grammar Enumeration and Inference”. Inf. and Control 26pp.236-255.
- [Ziv, 92]: J.Ziv and N.Merhav. : ”Estimating the Number of States of a Finite-State Source” IEEE Transactions on Information Theory Vol. 38 No.1 January 1992.