Grammatical Inference and Computational Linguistics

Menno van Zaanen
Tilburg Centre for Creative Computing
Tilburg University
Tilburg, The Netherlands
mvzaanen@uvt.nl

Colin de la Higuera
University of Saint-Étienne
France
cdlh@univ-st-etienne.fr

1 Grammatical inference and its links to natural language processing

When dealing with language, (machine) learning can take many different faces, of which the most important are those concerned with learning languages and grammars from data. Questions in this context have been at the intersection of the fields of inductive inference and computational linguistics for the past fifty years. To go back to the pioneering work, Chomsky (1955; 1957) and Solomonoff (1960; 1964) were interested, for very different reasons, in systems or programs that could deduce a language when presented information about it.

Gold (1967; 1978) proposed a little later a unifying paradigm called identification in the limit, and the term of grammatical inference seems to have appeared in Horning's PhD thesis (1969).

Out of the scope of linguistics, researchers and engineers dealing with pattern recognition, under the impulsion of Fu (1974; 1975), invented algorithms and studied subclasses of languages and grammars from the point of view of what could or could not be learned.

Researchers in machine learning tackled related problems (the most famous being that of inferring a deterministic finite automaton, given examples and counter-examples of strings). Angluin (1978; 1980; 1981; 1982; 1987) introduced the important setting of active learning, or learning for queries, whereas Pitt and his colleagues (1988; 1989; 1993) gave several complexity inspired results with which the hardness of the different learning problems was exposed.

Researchers working in more applied areas, such as computational biology, also deal with strings. A number of researchers from that field worked on learning grammars or automata from string data (Brazma and Cerans, 1994; Brazma, 1997; Brazma et al., 1998). Similarly, stemming from computational linguistics, one can point out the work relating language learning with more complex grammatical formalisms (Kanazawa, 1998), the more statistical approaches based on building language models (Goodman, 2001), or the different systems introduced to automatically build grammars from sentences (van Zaanen, 2000; Adriaans and Vervoort, 2002). Surveys of related work in specific fields can also be found (Natarajan, 1991; Kearns and Vazirani, 1994; Sakakibara, 1997; Adriaans and van Zaanen, 2004; de la Higuera, 2005; Wolf, 2006).

2 Meeting points between grammatical inference and natural language processing

Grammatical inference scientists belong to a number of larger communities: machine learning (with special emphasis on inductive inference), computational linguistics, pattern recognition (within the structural and syntactic sub-group). There is a specific conference called ICGI (*International Colloquium on Grammatical Inference*) devoted to the subject. These conferences have been held at Alicante (Carrasco and Oncina, 1994), Montpellier (Miclet and de la Higuera, 1996), Ames (Honavar and Slutski, 1998), Lisbon (de Oliveira, 2000), Amsterdam (Adriaans et al., 2002), Athens (Paliouras and Sakakibara, 2004), Tokyo (Sakakibara et al., 2006) and Saint-Malo (Clark et al., 2008). In the proceedings of this event it is possible to find a number of technical papers. Within this context, there has been a growing trend towards problems of language learning in the field of computational linguistics.

The formal objects in common between the two communities are the different types of automata and grammars. Therefore, another meeting point between these communities has been the different workshops, conferences and journals that focus on grammars and automata, for instance,

FSMNLP,GRAMMARS, CIAA, ...

3 Goal for the workshop

There has been growing interest over the last few years in learning grammars from natural language text (and structured or semi-structured text). The family of techniques enabling such learning is usually called "grammatical inference" or "grammar induction".

The field of grammatical inference is often subdivided into formal grammatical inference, where researchers aim to proof efficient learnability of classes of grammars, and empirical grammatical inference, where the aim is to learn structure from data. In this case the existence of an underlying grammar is just regarded as a hypothesis and what is sought is to better describe the language through some automatically learned rules.

Both formal and empirical grammatical inference have been linked with (computational) linguistics. Formal learnability of grammars has been used in discussions on how people learn language. Some people mention proofs of (non-)learnability of certain classes of grammars as arguments in the empiricist/nativist discussion. On the more practical side, empirical systems that learn grammars have been applied to natural language. Instead of proving whether classes of grammars can be learnt, the aim here is to provide practical learning systems that automatically introduce structure in language. Example fields where initial research has been done are syntactic parsing, morphological analysis of words, and bilingual modelling (or machine translation).

This workshop organized at EACL 2009 aimed to explore the state-of-the-art in these topics. In particular, we aimed at bringing formal and empirical grammatical inference researchers closer together with researchers in the field of computational linguistics.

The topics put forward were to cover research on all aspects of grammatical inference in relation to natural language (such as, syntax, semantics, morphology, phonology, phonetics), including, but not limited to

- Automatic grammar engineering, including, for example,
  - parser construction,
  - parameter estimation,
  - smoothing, ...
- Unsupervised parsing
- Language modelling
- Transducers, for instance, for
  - morphology,
  - text to speech,
  - automatic translation,
  - transliteration,
  - spelling correction, ...
- Learning syntax with semantics,
- Unsupervised or semi-supervised learning of linguistic knowledge,
- Learning (classes of) grammars (e.g. subclasses of the Chomsky Hierarchy) from linguistic inputs,
- Comparing learning results in different frameworks (e.g. membership vs. correction queries),
- Learning linguistic structures (e.g. phonological features, lexicon) from the acoustic signal,
- Grammars and finite state machines in machine translation,
- Learning setting of Chomskyan parameters,
- Cognitive aspects of grammar acquisition, covering, among others,
  - developmental trajectories as studied by psycholinguists working with children,
  - characteristics of child-directed speech as they are manifested in corpora such as CHILDES, ...
- (Unsupervised) Computational language acquisition (experimental or observational),

4 The papers

The workshop was glad to have as invited speaker Damir Ćavar, who presented a talk titled: *On bootstrapping of linguistic features for bootstrapping grammars*.

The papers submitted to the workshop and reviewed by at least three reviewers each, covered a very wide range of problems and techniques. Arranging them into patterns was not a simple task!

There were three papers focussing on transducers:

- Jeroen Geertzen shows in his paper *Dialogue Act Prediction Using Stochastic Context-Free Grammar Induction*, how grammar induction can be used in dialogue act prediction.

- In their paper (*Experiments Using OSTIA for a Language Production Task*), Dana Angluin and Leonor Becerra-Bonache build on previous work to see the transducer learning algorithm OSTIA as capable of translating syntax to semantics.

- In their paper titled *GREAT: a finite-state machine translation toolkit implementing a Grammatical Inference Approach for Transducer Inference (GIATI)*, Jorge González and Francisco Casacuberta build on a long history of GOATI learning and try to eliminate some of the limitations of previous work. The learning concerns finite-state transducers from parallel corpora.

Context-free grammars of different types were used for very different tasks:

- Alexander Clark, Remi Eyraud and Amaury Habrard (*A note on contextual binary feature grammars*) propose a formal study of a new formalism called "CBFG", describe the relationship of CBFG to other standard formalisms and its appropriateness for modelling natural language.

- In their work titled *Language models for contextual error detection and correction*, Herman Stehouwer and Menno van Zaanen look at spelling problems as a word prediction problem. The prediction needs a language model which is learnt.

- A formal study of French treebanks is made by Marie-Hélène Candito, Benoit Crabbé and Djamé Seddah in their work: *On statistical parsing of French with supervised and semi-supervised strategies.*

- Franco M. Luque and Gabriel Infante-Lopez study the learnability of NTS grammars with reference to the Penn treebank in their paper titled *Upper Bounds for Unsupervised Parsing with Unambiguous Non-Terminally Separated Grammars*.

One paper concentrated on morphology :

- In *A comparison of several learners for Boolean partitions: implications for morphological paradigm*, Katya Pertsova compares a rote learner to three morphological paradigm learners.

References

P. Adriaans and M. van Zaanen. 2004. Computational grammar induction for linguists. *Grammars*, 7:57–68.

P. Adriaans and M. Vervoort. 2002. The EMILE 4.1 grammar induction toolbox. In Adriaans et al. (Adriaans et al., 2002), pages 293–295.

P. Adriaans, H. Fernau, and M. van Zaannen, editors. 2002. *Grammatical Inference: Algorithms and Applications, Proceedings of* ICGI *'02*, volume 2484 of LNAI, Berlin, Heidelberg. Springer-Verlag.

D. Angluin. 1978. On the complexity of minimum inference of regular sets. *Information and Control*, 39:337–350.

D. Angluin. 1980. Inductive inference of formal languages from positive data. *Information and Control*, 45:117–135.

D. Angluin. 1981. A note on the number of queries needed to identify regular languages. *Information and Control*, 51:76–87.

D. Angluin. 1982. Inference of reversible languages. *Journal of the Association for Computing Machinery*, 29(3):741–765.

D. Angluin. 1987. Queries and concept learning. *Machine Learning Journal*, 2:319–342.

A. Brazma and K. Cerans. 1994. Efficient learning of regular expressions from good examples. In AII *'94: Proceedings of the 4th International Workshop on Analogical and Inductive Inference*, pages 76–90. Springer-Verlag.

A. Brazma, I. Jonassen, J. Vilo, and E. Ukkonen. 1998. Pattern discovery in biosequences. In Honavar and Slutski (Honavar and Slutski, 1998), pages 257–270.

A. Brazma, 1997. *Computational learning theory and natural learning systems*, volume 4, chapter Efficient learning of regular expressions from approximate examples, pages 351–366. MIT Press.

R. C. Carrasco and J. Oncina, editors. 1994. *Grammatical Inference and Applications, Proceedings of* ICGI *'94*, number 862 in LNAI, Berlin, Heidelberg. Springer-Verlag.

N. Chomsky. 1955. *The logical structure of linguistic theory*. Ph.D. thesis, Massachusetts Institute of Technology.

N. Chomsky. 1957. *Syntactic structure*. Mouton.

A. Clark, F. Coste, and L. Miclet, editors. 2008. *Grammatical Inference: Algorithms and Applications, Proceedings of* ICGI *'08*, volume 5278 of LNCS. Springer-Verlag.

C. de la Higuera. 2005. A bibliographical study of grammatical inference. *Pattern Recognition*, 38:1332–1348.

A. L. de Oliveira, editor. 2000. *Grammatical Inference: Algorithms and Applications, Proceedings of* ICGI *'00*, volume 1891 of LNAI, Berlin, Heidelberg. Springer-Verlag.

K. S. Fu and T. L. Booth. 1975. Grammatical inference: Introduction and survey. Part I and II. IEEE *Transactions on Syst. Man. and Cybern.*, 5:59–72 and 409–423.

K. S. Fu. 1974. *Syntactic Methods in Pattern Recognition*. Academic Press, New-York.

E. M. Gold. 1967. Language identification in the limit. *Information and Control*, 10(5):447–474.

E. M. Gold. 1978. Complexity of automaton identification from given data. *Information and Control*, 37:302–320.

J. Goodman. 2001. A bit of progress in language modeling. Technical report, Microsoft Research.

V. Honavar and G. Slutski, editors. 1998. *Grammatical Inference, Proceedings of* ICGI *'98*, number 1433 in LNAI, Berlin, Heidelberg. Springer-Verlag.

J. J. Horning. 1969. *A study of Grammatical Inference*. Ph.D. thesis, Stanford University.

M. Kanazawa. 1998. *Learnable Classes of Categorial Grammars*. CSLI Publications, Stanford, Ca.

M. J. Kearns and U. Vazirani. 1994. *An Introduction to Computational Learning Theory*. MIT press.

L. Miclet and C. de la Higuera, editors. 1996. *Proceedings of* ICGI *'96*, number 1147 in LNAI, Berlin, Heidelberg. Springer-Verlag.

B. L. Natarajan. 1991. *Machine Learning: a Theoretical Approach*. Morgan Kauffman Pub., San Mateo, CA.

G. Paliouras and Y. Sakakibara, editors. 2004. *Grammatical Inference: Algorithms and Applications, Proceedings of* ICGI *'04*, volume 3264 of LNAI, Berlin, Heidelberg. Springer-Verlag.

L. Pitt and M. Warmuth. 1988. Reductions among prediction problems: on the difficulty of predicting automata. In *3rd Conference on Structure in Complexity Theory*, pages 60–69.

L. Pitt and M. Warmuth. 1993. The minimum consistent DFA problem cannot be approximated within any polynomial. *Journal of the Association for Computing Machinery*, 40(1):95–142.

L. Pitt. 1989. Inductive inference, DFA's, and computational complexity. In *Analogical and Inductive Inference*, number 397 in LNAI, pages 18–44. Springer-Verlag, Berlin, Heidelberg.

Y. Sakakibara, S. Kobayashi, K. Sato, T. Nishino, and E. Tomita, editors. 2006. *Grammatical Inference: Algorithms and Applications, Proceedings of* ICGI *'06*, volume 4201 of LNAI, Berlin, Heidelberg. Springer-Verlag.

Y. Sakakibara. 1997. Recent advances of grammatical inference. *Theoretical Computer Science*, 185:15–45.

R. Solomonoff. 1960. A preliminary report on a general theory of inductive inference. Technical Report ZTB-138, Zator Company, Cambridge, Mass.

R. Solomonoff. 1964. A formal theory of inductive inference. *Information and Control*, 7(1):1–22 and 224–254.

M. van Zaanen. 2000. ABL: Alignment-based learning. In *Proceedings of* COLING *2000*, pages 961–967. Morgan Kaufmann.

G. Wolf. 2006. *Unifying computing and cognition*. Cognition research.