Bio-inspired Grammatical Inference

Leonor Becerra-Bonache

Bio-inspired Grammatical Inference

Leonor Becerra-Bonache

Laboratoire Hubert Curien, UMR CNRS 5516
Université de Saint-Etienne, Jean Monnet
Rue du Professeur Benoit Lauras, 42000 Saint-Etienne, France
leonor.becerra@univ-st-etienne.fr

Abstract. The field of Grammatical Inference was originally motivated by the problem of natural language acquisition. However, the formal models proposed within this field have left aside this linguistic motivation. In this paper, we propose to improve models and techniques used in Grammatical Inference by using ideas coming from linguistic studies. In that way, we try to give a new bio-inspiration to this field.

1 Introduction

The problem of how children acquire their first language has attracted the attention of researchers for many years. The desire to better understand natural language acquisition has motivated research in formal models of language learning [27,26]. Such models are of great interest for several reasons. On one hand, these models can help us to answer several key questions about natural language learning. On the other hand, these formal models can provide an operational framework for practical applications of language learning; for example, language learning by machines.

Grammatical Inference (GI) is a subfield of Machine Learning that deals with the learning of formal languages. The initial theoretical foundations of GI were given by E.M. Gold [18], who tried to formalize the process of natural language acquisition. After Gold's seminal work, research in this field has been specially focused on obtaining formal results (e.g., formal descriptions of the languages to be learned, formal proofs that a concrete algorithm can efficiently learn according to some concrete denitions, etc) [13]. Several formal models of language learning have been proposed in this field [17], however, such models do not take into account some important aspects of natural language acquisition, and assume idealized conditions as compared to the conditions under which children learn language (as we will see in Section 2). Therefore, although GI studies were motivated by the problem of natural language acquisition, its mathematization has left natural approaches aside.

Since the study of formal models of language learning is of great interest to better understand natural language acquisition, it is important that such models are inspired by studies of natural language acquisition. In that way, models can be more realistic, and can better simulate the human processing and acquisition of language.

J.M. Ferrández et al. (Eds.): IWINAC 2011, Part I, LNCS 6686, pp. 313–322, 2011.

Based on all these ideas, we propose to use ideas coming from linguistic studies to improve the models and techniques used in GI. In that way, we propose to give a new bio-inspiration to the field of GI, and bring back Grammatical Inference to its origins.

2 Grammatical Inference Studies

Grammatical Inference deals with the learning of formal languages from a set of data. The basic framework can be considered as a game played between two players: a teacher and a learner. The teacher provides information to the learner, and the learner must identify the underlying language from that information [13]. Excellent surveys on this field can be found in [28,16,17].

As we can see, the process of learning a formal language has some similarities with the process of language acquisition. For example, instead of a teacher and a learner, we have an adult and a child. Moreover, a child learns a language from the data that he/she receives (a child with an English environment will learn to speak English, and the same child with a Japanese environment will learn to speak Japanese). Therefore, GI provides a *good theoretical framework* for investigating the process of language learning.

The relevance of formal results in GI to the question of how children acquire their native language has been well recognized [33,13].

> *Positive results can help us to understand how humans might learn languages by outlining the class of algorithms that might be used by humans, considered as computational systems at a suitable abstract level. Conversely, negative results might be helpful if they could demonstrate that no algorithms of a certain class could perform the task in this case we could know that the human child learns his language in some other way* [13, p. 26].

However, work in the field of GI has been specially focused on the mathematical aspects of the language learning problem, leaving aside the linguistic motivation that originated these studies. Next we review some of the main drawbacks of GI from a linguistic point of view.

2.1 Formal Models in Grammatical Inference

Several formal models of language learning have been proposed in the field of GI. The main ones are: *Identification in the limit* [18], *Query learning model* [1], and *PAC learning model* [32]. The main drawback of these models is that they do not take into account some relevant aspects of natural language learning.

In the model proposed by E.M. Gold, *Identification in the limit*, there is no limit on how long the learner can take to guess the correct language; from a linguistic point of view, efficiency is important, since children are able to learn the language in an efficient way. Moreover, the learner passively receives strings of the language (but natural language learning is more than that, children also interact

with their environment) and hypothesizes complete grammars instantaneously (this assumption is unrealistic).

The *Query learning model* proposed by D. Angluin has also some controversial aspects from a linguistic point of view; for example, the learner is able to ask the teacher if his hypothesis is correct (such a query will never be produced in a real situation; a child would never ask the adult if his grammar is the correct one), and the learner learns exactly the target language (this is not realistic, since everybody has imperfections in their linguistic competence).

In the *PAC learning model* proposed by L. Valiant, the examples provided to the learner have the same distribution throughout the process; this requirement is too strong for practical applications.

Therefore, none of these models perfectly account for natural language acquisition. Research in GI has been focused on the mathematical aspects of the formal models proposed, without exploiting their linguistic relevance. A longer discussion about these models can be found in [5].

2.2 Language Learning Problem

The problem of language learning concerns both the acquisition of the *syntax* (i.e., rules for generating and recognizing correct sentences in the language) and the *semantics* (i.e., the underlying meaning of each sentence) of a target language [26]. However, GI studies has been focused only on learning the syntax.

Semantics not only is one component of language learning, but also seems to play an important role in the first stages of children's language acquisition (as we will see in the next section). Therefore, it is also of great interest to study this component. Unfortunately, all these considerations have not been taken into account in GI studies; the learning problem has been reduced to syntax learning, and all semantic information has been omitted from their works.

GI algorithms are based on the availability of different types of information: positive examples, negative examples, the presence of a teacher able to answer queries, etc. However, *what kind of data is available to children?* Ideally, to better understand the process of natural language acquisition and to correctly simulate it, we should provide to our algorithm the same kind of examples that are available to children. However, some of the data used by GI algorithms are controversial from a linguistic point of view. We will discuss some linguistic studies that try to answer this question in the next section.

In order to make the problem of language learning well defined, it is also necessary to choose an appropriate class of grammars. The classes of *regular* and *context-free* grammars are often used in GI to model the target grammar. These two classes constitutes the first two levels of the Chomsky hierarchy. Thus, the following question arises: do they have enough expressive power to describe natural languages? From a linguistic point of view, it is of great interest to study classes of grammars that are able to generate the most relevant constructions that appear in natural languages. However, it seems not to be the case of regular and context-free grammars. We will discuss the limitations of the Chomsky hierarchy in the next section.

3 Linguistic Studies

The question of how children acquire their native language has been traditionally addressed by linguists. Their approach is specially focused on making experiments with children that are learning their native language. In that way, they try to collect data about the process of natural language acquisition (e.g., first sentences produced by children, errors made, etc.). Their final goal is to investigate the mental processes that occur during children's language acquisition, and try to describe this process.

There are different types of experiments. Depending on the way in which the data is collected, we have a *naturalistic approach* (i.e., samples of child language is collected or recorded in a comfortable environment. Data is collected regularly) or a *experimental approach* (i.e., the researcher proposes a work hypothesis and design specific tasks that have to be performed by the child to use specific language structures. A statistical analysis of the data is done at the end). Depending on the number of children used to do the experiments, we have *longitudinal studies* (i.e., experiments are focused just on one child, and they are done over a long period of time. Such approach is often combined with the naturalistic approach) or *transversal studies* (experiments are made with a group of children of different ages. Such approach is often combined with an experimental approach). And finally, depending on the kind of tasks performed by the child, we have experiments based on comprehension, production or imitation.

The CHILDES database (Child Language Data Exchange system) provides a large amount of useful data for linguistic studies of children's language acquisition. In this database we can find transcript and media data collected from conversations between children and adults. It has content in over 20 languages from 130 different corpora, all of which are available in `http://childes.psy.cmu.edu/`

Studies carried out in the field of Linguistics have helped to better understand some aspects of natural language acquisition [14]. However, there is still a lot of questions that do not have a clear answer; for example, what factors really have an effect in the process of children's language acquisition. Therefore, despite all investigations conducted so far, it has not been possible yet to understand all the rules, strategies, and other processes that underlie children's language acquisition.

Next we review some of the works and results obtained in this field, that are relevant for GI studies.

3.1 Data Available to Children

The question of what kind of data is available to children during the learning process is still a subject of discussion in Linguistics. It is generally accepted that children receive sentences that are grammatically correct, that is, *positive data*. However, the availability of another kind of data, usually called *negative data*, is a matter of controversy.

There are three different proposals to this question. The first proposal is that children do not receive negative data and they must rely on innate information to

acquire their native language. This proposal is based on the Chomsky's *poverty of stimulus* argument: there are principles of the grammar that cannot be learnt from only positive data, and since children do not receive negative data (i.e., evidence about what is not grammatical), one can conclude that the innate linguistic capacity is what provides the additional knowledge that is necessary for language learning. Moreover, work presented by Brown and Hanlon in [11], has been used as an argument to support the unavailability of negative data to children. Concretely, they analyzed adult approval and disapproval of child utterances (for example, adult's answers as "That's right", "Correct", "That's wrong", "No"), and they found no relation between this type of answers and the grammaticality of the sentences produced by the children. However, this approach raises several questions: Should only explicit disapproval count as negative evidence? Could adults correct children in a different way?

The second proposal is that children receive negative data in the form of *different reply-types* given in response to grammatical versus ungrammatical child utterances. Hirsh-Pasek et al. [19] and Morgan and Travis [24] studied this type of negative evidence and proposed that parents respond to ungrammatical child utterances by using different types of answers from those they use when responding to grammatical utterances. Under this view, the reply-type would indicate to the child whether an utterance was grammatically correct or not. The problem of this second approach is that it does not take into account whether the adult's replies contain corrective information [12]. Moreover, under this approach, children would learn what utterances are correct only after complex statistical comparisons [23].

The third proposal is that children receive negative evidence in the form of *reformulations*, and they not only can detect them, but also they can make use of that information. Reformulations are sentences adults use in checking up on what their children intended to say (for example, a child says "milk milk" and the father answers "you want milk"?). Chouinard and Clark [12] proposed this new view of negative evidence. The main properties of this kind of corrections are the following: i) Adult's correction preserves the same meaning of the child; ii) Adult uses the correction to keep the conversation on track (adult reformulates the sentence just to make sure that he has understood the child's intentions); iii) Child's utterance and adult's correction have the same meaning, but different form.

It is worth noting that reformulations are often provided to children during the early stages of children's language acquisition. Moreover, semantics seems to play an important role in the first stages of children's language acquisition, concretely in the stage known as the two-word stage, in which children go through the production of one word to the combination of two elements [29,30].

3.2 Location of Natural Languages in the Chomsky Hierarchy

The question of where natural languages are located in the Chomsky hierarchy has been a subject of debate for a long time. This question was posed by Chomsky in the 50's. The debate was focused on whether natural language are context-free or not.

There were many attempts to prove the non-context-freeness of natural languages, but we have to wait until the late 80's to find solid arguments that support this idea. By that time, Bresnan et al. [10], Culy [15] and Shieber [31] presented some clear examples of natural language structures that cannot be described using a context-free grammar. Such examples were found in three different natural languages: Dutch, Bambara and Swiss-German. The kind of structures founded in these languages are: multiple agreements, crossed agreements and duplication. Therefore, after 20 years, linguists seemed to agree that context-free languages do not have enough expressiveness to describe the main context-sensitive constructions found in natural languages.

The discovery of non-context-free structures in natural languages aroused out the study and development of grammatical formalisms with more generative power than context-free. Since context-sensitive seems not to be the good solution (they are too powerful and computationally too complex), researchers tried to find another formal grammar more adequate to model natural language structures. The idea of generating context-free and non-context-free structures, keeping under control the generative power, has led to the notion of *Mildly Context-Sensitive* (MCS) grammars [20].

There exist very well known mechanisms to fabricate MCS families, for example, tree adjoining grammars, head grammars, combinatory categorial grammars. All of them occupy a concentric position in the Chomsky hierarchy, between context-free and context-sensitive. However, is it necessary that such formalism generates all context-free languages?

As some authors point out [21,22], natural languages could occupy an orthogonal position in the Chomsky hierarchy, that is, they contain some regular languages and some context-free languages, but they are included in context-sensitive. In fact, we can find natural language constructions that are neither regular nor context-free, and also some regular or context-free constructions that do not appear naturally in sentences.

4 Bio-inspired Grammatical Inference

In Section 2 we have discussed some of the main drawbacks of GI from a linguistic point of view. Concretely, we have seen that the most important models studied in GI do not take into account some important aspects of children's language acquisition and, consequently, they are quite unrealistic. The data used by the GI algorithms are also controversial from a linguistic point of view. Moreover, we have pointed out that works in GI tend to omit the semantic information and reduce the learning problem to syntax learning. And finally, we have seen that research in GI have been focused on learning regular and context-free languages, which constitutes the classes with less generative power of the Chomsky hierarchy.

In Section 3 we have discussed some of the works and results obtained in Linguistics, concerning natural language acquisition. Concretely, we have seen that there are three different proposals about the type of data that is available

to the child. We have also pointed out the relevance of semantics in the first stages of children's language acquisition. And finally, we have seen that the Chomsky hierarchy has some limitations, specially when we try to locate natural languages in this hierarchy; since regular and context-free grammars seems not to be very adequate to model natural language syntax, linguists have tried to find other grammatical formalisms that have interesting linguistic and computational properties, such as Mildly Context-Sensitive.

Taking into account all these ideas, we propose to use linguistic studies to improve models and techniques used in GI. Thanks to ideas coming from linguistic studies on natural language acquisition, models in GI could be more realistic; these models could take into account more aspects about children's language acquisition. Moreover, such ideas could also improve the results obtained in the field of GI. In that way, we would use a *bio-inspired* model for language learning.

First of all, we propose that GI algorithms take into account not only *positive data*, but also *corrections* during the learning process. We consider that the most convincing proposal to the question of what kind of data is available to children, is the one proposed by Chouinard and Clark [12]. To consider that only explicit disapproval counts as negative data is not realistic. As Chouinard and Clark showed, adults correct children in a very different way, taking into account the meaning that the child intends to express. Moreover, a very large number of examples of such kind of meaning-preserving corrections can be found in real conversations between children and adults (for example, in CHILDES database). The second proposal is also unconvincing, since corrective and non-corrective replies are mixed in their analysis, and hence, learning from "reply-types" would require that children do complex statistical comparisons in order to learn which sentences are correct. Therefore, as Chouinard and Clark proposed in [12], we consider that meaning-preserving corrections are available to children, and they can help them to learn some aspects of natural language syntax.

In order to see the effect of corrections in language leaning, we propose to incorporate the idea of corrections to the studies of GI. We have already done some work in this direction. Our first approach has consisted on considering only *syntactic corrections* based on proximity between strings. Since this idea was totally new in GI, we started by learning deterministic finite automata [8], in the framework of query learning (i.e., the learner is able to ask queries to the teacher, and the teacher has to answer correctly to these questions). Later, these results were extended to learn other classes with interesting properties, such as *balls of strings* (which are defined by using the edit distance) [7]. In both cases, when the learner asks for a string that does not belong to the target language, the teacher returns a correction (in the first case, such correction is based on the shortest extension of the queried string, and in the second case, such correction is based on the edit distance). In both cases, we could show that results can be improved thanks to corrections. Therefore, such works show that new challenging results can be obtained in the field of GI if corrections are taken into account in the learning process.

Our second approach is based on *semantic corrections*. Corrections have a semantic component that has not been taken into account in previous works. Hence, we have proposed a new computational model of language learning that takes into account semantics. This model is bio-inspired by studies on children's language acquisition, and more concretely, by the results obtained by Chouinard and Clark in [12]. Our final goal has been to find a formal model that gives an account of *meaning preserving corrections*, and in which we can address the following questions: What are the effects of corrections on learning syntax? Can corrections facilitate the language learning process? Can semantic information simplify the problem of learning formal languages? It is worth noting that this has been the first attempt to incorporate a robust notion of semantics in the field of GI. Such a model has allowed us to investigate aspects of the roles of semantics and corrections in the process of learning to understand and speak a natural language. Our main results can be found in [4,3,2].

Taking into account that GI studies have been focused on learning regular and context-free languages, but, as linguistic studies suggest, these classes have limited expressive power to describe natural language syntax, we propose that GI studies focus on other classes such as *Mildly context-sensitive*. Moreover, we also support the idea that natural languages occupy an orthogonal position in the Chomsky hierarchy. Therefore, we propose to study formalisms that are able to generate MCS languages (i.e., they generate multiple agreement, crossed agreement and duplication structures, and they are computational feasible), and that occupy an orthogonal position in the Chomsky hierarchy (i.e., they contain some regular, some context-free, but they are included in context-sensitive). We have also done some works in this direction. We studied a mechanism that have such interesting properties, called *Simple External Contextual*. Our main results can be found in [9,25,6].

5 Conclusions

The field of GI provides a good theoretical framework for investigating the process of natural language acquisition. However, studies in this field have been focused on the mathematical aspects of the formal models proposed, without exploiting their linguistic relevance. Therefore, the linguistic motivation that originated GI studies has been left aside.

In this paper, we have discussed some linguistic studies on children's language acquisition and we have proposed to use them in order to improve models and techniques used in GI. Concretely, we have proposed that GI studies take into account *corrections* and *semantics* during the learning process, and they focus on classes of languages that are MCS and occupy an orthogonal position in the Chomsky hierarchy.

We have also present some works in this line. These works show that new challenging results in the field of GI can be obtained. Moreover, models in GI can be improved by using these linguistics ideas.

It is worth noting that a formal model bio-inspired by all these linguistic ideas could also help us to better understand natural language acquisition. By

studying formal models of language learning, several key questions in linguistics can be answered, as for example, the type of input available to the learner, the impact of semantic information on learning the syntax of a language, etc. In fact, the model that we have proposed in [2] tries to answer some of these questions.

References

1. Angluin, D.: Learning regular sets from queries and counterexamples. Information and Computation 75, 87–106 (1987)
2. Angluin, D., Becerra-Bonache, L.: A model of semantics and corrections in language learning. In: YALEU/DCS/TR-1425 (April 2010)
3. Angluin, D., Becerra Bonache, L.: Experiments with an algorithm to learn meaning before syntax. In: ForLing2008, pp. 1–12 (2008)
4. Angluin, D., Becerra-Bonache, L.: Learning meaning before syntax. In: Clark, A., Coste, F., Miclet, L. (eds.) ICGI 2008. LNCS (LNAI), vol. 5278, pp. 1–14. Springer, Heidelberg (2008)
5. Becerra-Bonache, L.: On the Learnability of Mildly Context-Sensitive Languages using Positive Data and Correction Queries. PhD thesis, Rovira i Virgili University (2006)
6. Becerra-Bonache, L., Case, J., Jain, S., Stephan, F.: Iterative learning of simple external contextual languages. In: Freund, Y., Györfi, L., Turán, G., Zeugmann, T. (eds.) ALT 2008. LNCS (LNAI), vol. 5254, pp. 359–373. Springer, Heidelberg (2008)
7. Becerra-Bonache, L., de la Higuera, C., Janodet, J.C., Tantini, F.: Learning balls of strings from edit corrections. Journal of Machine Learning Research 9, 1841–1870 (2008)
8. Becerra-Bonache, L., Dediu, A.-H., Tîrnăucă, C.: Learning DFA from correction and equivalence queries. In: Sakakibara, Y., Kobayashi, S., Sato, K., Nishino, T., Tomita, E. (eds.) ICGI 2006. LNCS (LNAI), vol. 4201, pp. 281–292. Springer, Heidelberg (2006)
9. Becerra-Bonache, L., Yokomori, T.: Learning mild context-sensitiveness: Toward understanding children's language learning. In: Paliouras, G., Sakakibara, Y. (eds.) ICGI 2004. LNCS (LNAI), vol. 3264, pp. 53–64. Springer, Heidelberg (2004)
10. Bresnan, J., Kaplan, R.M., Peters, S., Zaenen, A.: Cross-serial dependencies in dutch. In: Savitch, W.J., Bach, E., Marsh, W., Safran-Naveh, G. (eds.) The Formal Complexity of Natural Language, pp. 286–319. D. Reidel, Dordrecht (1987)
11. Brown, R., Hanlon, C.: Derivational complexity and order of acquisition in child speech. In: Hayes, J.R. (ed.) Cognition and the Development of Language, pp. 11–54. Wiley, New York (1970)
12. Chouinard, M.M., Clark, E.V.: Adult reformulations of child errors as negative evidence. Journal of Child Language 30, 637–669 (2003)
13. Clark, A.: Grammatical inference and first language acquisition. In: Psychocomputational Models of Human Language Acquisition, Geneva, pp. 25–32 (2004)
14. Clark, E.V.: First Language Acquistion. Cambridge University Press, Cambridge (2002)
15. Culy, C.: The complexity of the vocabulary of bambara. In: Savitch, W.J., Bach, E., Marsh, W., Safran-Naveh, G. (eds.) The Formal Complexity of Natural Language, pp. 349–357 (1987)
16. de la Higuera, C.: A bibliographical study of grammatical inference. Pattern Recognition 38, 1332–1348 (2005)

17. de la Higuera, C.: Grammatical inference: learning automata and grammars. Cambridge University Press, Cambridge (2010)
18. Gold, E.M.: Language identification in the limit. Information and Control 10, 447–474 (1967)
19. Hirsh-Pasek, K., Treiman, R.A., Schneiderman, M.: Brown and hanlon revisited: mothers sensitivity to ungrammatical forms. Journal of Child Language 11, 81–88 (1984)
20. Joshi, A.K.: How much context-sensitivity is required to provide reasonable structural descriptions: Tree adjoining grammars. In: Dowty, D., Karttunen, L., Zwicky, A. (eds.) Natural Language Parsing: Psychological, Computational and Theoretical Perspectives, pp. 206–250. Cambridge University Press, Cambridge (1985)
21. Kudlek, M., Martín-Vide, C., Mateescu, A., Mitrana, V.: Contexts and the concept of mild context-sensitivity. Linguistics and Philosophy 26(6), 703–725 (2002)
22. Manaster-Ramer, A.: Some uses and abuses of mathematics in linguistics. In: Martín-Vide, C. (ed.) Issues in Mathematical Linguistics, pp. 73–130. John Benjamins, Amsterdam (1999)
23. Marcus, G.F.: Negative evidence in language acquisition. Cognition 46, 53–95 (1993)
24. Morgan, J.L., Travis, L.L.: Limits on negative information in language input. Journal of Child Language 16, 531–552 (1989)
25. Oates, T., Armstrong, T., Bonache, L.B., Atamas, M.: Inferring grammars for mildly context sensitive languages in polynomial-time. In: Sakakibara, Y., Kobayashi, S., Sato, K., Nishino, T., Tomita, E. (eds.) ICGI 2006. LNCS (LNAI), vol. 4201, pp. 137–147. Springer, Heidelberg (2006)
26. Parekh, R., Honavar, V.: Grammar inference, automata induction and language acquisition. In: Moisl Dale and Somers, editors, pp. 727–774. Marcel Dekker, New York (2000)
27. Pinker, S.: Formal models of language learning. Cognition 7, 217–283 (1979)
28. Sakakibara, Y.: Recent advances of grammatical inference. Theoretical Computer Science 185, 15–45 (1997)
29. Schaerlaekens, A.M.: The two-word sentence in child language development. In: Mouton, The Hague (1973)
30. Schlesinger, I.M.: Production of utterances and language acquisition. In: Slobin, D.I. (ed.) The Ontogenesis of Grammar, pp. 63–103. Academic Press, New York-London (1971)
31. Shieber, S.M.: Evidence against the context-freeness of natural languages. In: Savitch, W.J., Bach, E., Marsh, W., Safran-Naveh, G. (eds.) The Formal Complexity of Natural Language, pp. 320–334. D. Reidel, Dordrecht (1987)
32. Valiant, L.G.: A theory of the learnable. Communication of the ACM 27, 1134–1142 (1984)
33. Wexler, K., Culicover, P.: Formal Principles of Languages Acquisition. MIT Press, Cambridge (1980)