Children as Models for Computers: Natural Language Acquisition for Machine Learning

Leonor Becerra-Bonache[1] and M. Dolores Jiménez-López[2]

[1]Laboratoire Hubert Curien, UMR CNRS 5516, Université de Saint-Etienne, Jean Monnet
Rue du Professeur Benoit Lauras, 42000 Saint-Etienne, France

[2]Research Group on Mathematical Linguistics, Universitat Rovira i Virgili
Av. Catalunya 35, 43005 Tarragona, Spain

Abstract. This paper focuses on a subfield of machine learning, the so-called grammatical inference. Roughly speaking, grammatical inference deals with the problem of inferring a grammar that generates a given set of sample sentences in some manner that is supposed to be realized by some inference algorithm. We discuss how the analysis and formalization of the main features of the process of human natural language acquisition may improve results in the area of grammatical inference.

1 Introduction

In the so-called information society there is a need for a comprehensive language technology for information management. While computers can communicate only through artificial languages designed specifically for them, its use will be restricted to a minority of people. The natural thing would be to allow users to speak to the computer in their own natural language. To solve the problem of communication between machines and humans it is necessary to construct artificial mechanisms to simulate the human processing and acquisition/learning of language. The computational models of language that have been proposed up to now are far from satisfactory. To reach suitable models is a problem that must be approached from an interdisciplinary perspective. In this interdisciplinary task, linguistics and the knowledge of how natural language is acquired and processed have a key role.

Artificial Intelligence aims to study and design intelligent machines. This field was founded in 1956. AI founders were very optimistic about the future of this new field. For example, H. Simon predicted that "machines will be capable, within twenty years, of doing any work a man can do" [7]. In 2010, we can state that this prediction has not come true yet. However, we have machines that are able to do "some of the things" that a man can do; for example, we have machines that are able to play soccer, to play some instruments, to express feelings by moving their face (e.g., MDX, KISMET), etc. Nevertheless, what has not been achieved yet is that machines learn to speak.

It is a truism that natural languages are very complex, but despite this complexity a child is able to efficiently learn a natural language in a very short time. Children are able

Becerra-Bonache L. and Dolores Jiménez-López M..
Children as Models for Computers: Natural Language Acquisition for Machine Learning.
DOI: 10.5220/0003309300670076
In *Proceedings of the 1st International Workshop on AI Methods for Interdisciplinary Research in Language and Biology* (BILC-2011), pages 67-76
ISBN: 978-989-8425-42-3

to learn any natural language given the adequate input, and they do so effortlessly, with a limited exposure to data, and without any specific training. Therefore, if we are able to give machines the capacity of learning language as children do, maybe we could reach to have computers that learn to speak. A computational simulation of natural language acquisition could be used to develop computer systems that can recognize, understand and generate natural languages, solving in this way the problem of communication between machines and humans.

Machine Learning is a field of Artificial Intelligence that aims to develop techniques that allow computers to learn. Concretely, it consists in designing and developing algorithms that allow to computers to change their behaviour based on some data. Grammatical Inference (GI) is a specialized subfield of Machine Learning that deals with the learning of formal languages from a set of data. To solve a GI problem requires, on one hand, a teacher that provides data to a learner, and on the other hand, a learner (or learning algorithm) that from that data must identify the underlying language. As we can see, this process has some similarities with the process of language acquisition (instead of a teacher and a learner, we have an adult and a child). Therefore, GI provides a good theoretical framework to investigate the idea of simulating some of the features of natural language acquisition in order to check if they could simplify or improve the problem of learning a language.

In general it is claimed that machine learning or GI can provide natural language processing/acquistion a range of alternative learning algorithms as well as additional general approaches and methodologies [1, 2]. So, it is accepted that GI models can help in the understanding of how humans process and acquire language. In this paper we claim the opposite direction, that is, that simulation of the process of acquiring a natural language could improve GI techniques and this improvement could have important implications in the field of human language technologies. Therefore, what we defend here is that natural language acquisition can help GI. If computers are able to learn a language like a human, they could use language like a human. If we are able to create machines with human-like capabilities (like learning a language), we will make possible for the user to interact with the computer, without any special skill or training, just as they would do to a person.

2 Grammatical Inference Models

The research field known as GI deals with the learning of formal languages. Roughly speaking, a GI problem 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 [2]. The initial theoretical foundations of GI were given by E.M. Gold [3]. A remarkable amount of research has been done after Gold's seminal work. Three formal models have been widely investigated in the field of GI:

- Identification in the limit [3].
- Query learning model [4].
- PAC learning model [5].

Each of these models is based on different learning settings (what kind of data is used in the learning process and how these data are provided to the learner) and different criteria for a successful inference (under what conditions we say that a learner has been successful in the language learning task).

In 1967 Gold introduced the first model for GI: *identification in the limit*. In this model, after each new example received, the learner (inference algorithm) must return some hypotheses. If the learner returns a correct answer and does not change its guess after this, then we can say that the identification is achieved. There are two traditional settings: i) Learning from text: only positive data (i.e. strings that belong to the languages to be learned) are given to the learner; ii) Learning from informant: positive and negative data (i.e., strings that do not belong to the language) are available to the learner.

A different learning paradigm that has been exhaustively studied in GI is *learning from queries*, introduced by Angluin. In the query learning model, there is a teacher (oracle) that knows the language and has to answer correctly specific kind of queries asked by the learner. Different kind of queries could be available to the learner, but membership queries (MQs) and equivalence queries (EQs) have established themselves as the standard combination to be used. In the case of a MQ, the learner asks if a string is in the language, and the teacher answers "yes" or "no". When the learner asks an EQ, he makes a conjecture and the teacher answers "yes" if it generates the same language, and if the answer is "no", a counterexample is returned.

Valiant introduced *probably approximately correct learning* (PAC learning) [5], which is a distribution-independent probabilistic model of learning from random examples. In this model, the inference algorithm takes a sample as input and produces a grammar as output. A successful inference algorithm is one that with high probability finds a grammar whose error is small. In this PAC learning model, more negative results have been proved than positive results (for GI). Even for the case of DFA, most results are negative. The requirement that the learning algorithm must learn under any arbitrary (but fixed) probability distribution seems too strong.

Each of these models have aspects that make them useful to study the problem of natural language acquisition to a certain extent, but other aspects of the models make them unsuitable for this task. For example, in Gold's model, there is not limit on how long it can take the learner to guess the correct language (but children are able to learn language in an efficient way), the learner hypothesizes complete grammars instantaneously (this is not the case in children's language acquisition), and the learner passively receives strings of the language (but children also interact with their environment). In Angluin's model, the queries introduced in this model are quite unnatural for real learning environments (a child will never ask if his/her grammar is the correct one). Moreover, the learner has to learn exactly the target language (but everybody has imperfections in their linguistic competence) and the teacher is assumed to be perfect (i.e., he knows everything and always gives the correct answers. This is an ideal teacher that does not occur in a real situation). In Valiant's model, the requirement that the examples have the same distribution throughout the process is too strong for practical situations. Therefore, none of these models perfectly accounts for natural language acquisition.

3 Grammatical Inference as Natural Language Acquistion

The problem of language learning in GI presents similarities with the process of language acquisition. In a GI problem we have a teacher and a learner, the teacher provides information about a language to the learner and the learner must infer the grammar for that language. Similarly, in a process of natural language acquisition, the child receives information from the adult and has to learn the grammar under that language. Taking into account this similarities, and the fact that, despite the complexity of language, a child is able to efficiently learn a natural language in a very short time, what we propose in this section is to implement in GI algorithms some of the main characteristics of natural language acquisition, in order to check if they could simplify or improve the problem of learning a language in the field of GI.

We deal with three different aspects. First, we discuss about the class of languages to be learned. Second, we take into account the type of data/information that should be provided to the algorithm. And finally, we discuss which component of the grammar should be learned.

4 The Language to be Learned

An important question in GI models is the type of grammars that must be learned. Most of them try to learn regular and context-free grammars and languages. However, limitations of the Chomsky hierarchy to describe natural languages are well known.

The question whether grammatical sentences of natural languages form regular, context-free, context-sensitive or recursively enumerable sets has been subject to many discussions since it was posed by Chomsky in 1957. There seems to be little agreement among linguists concerning the position of natural languages in the Chomsky hierarchy. It seems that neither the family of regular or context-free languages have enough expressiveness to describe the basic context-sensitive syntactic constructions found in natural languages. Several attempts have been made to prove the non-context-freeness of natural languages [6, 7]. Despite the fact that the non-context-freeness of natural language has become the standardly accepted theory, there are linguists such as Pullum and Gazdar who, after reviewing the various attempts to establish that natural languages are not context-free, come to the conclusion that every published argument purporting to demonstrate the non-context-freeness of some natural language is invalid, either formally or empirically or both [8]. Despite these arguments, it seems to be an untenable position that all syntactical aspects in natural languages can be captured by context-free grammars. However, the overwhelming bulk of natural language syntax is context-sensitive. Therefore, it is of interest to study grammatical formalisms with more generative power than CF. However, context-sensitive grammars seems not to be the right solution: they are too powerful, many problems are undecidable, etc. Therefore, it is desirable to find intermediate generative devices able of conjoining the simplicity of context-free grammars with the power of context-sensitive ones.

Within the field of formal languages, the above idea has led to the branch of *Regulated Rewriting* [9]. *Matrix grammars*, *programmed and controlled grammars*, *random*

context grammars, *conditional grammars*, etc. are examples of devices that use context-free grammars while applying some restrictions to the rewriting process in order to obtain context-free structures as well as the non-context-free constructions present in natural language. However, those devices present, in general, an excesive big generative power that leads to the generation of structures non-significative for natural languages. The idea of keeping under control the generative power, while generating context-free structures and non-context-free constructions, has led to the so-called *mildly context-sensitive grammars* [10]. *Tree adjoining-grammars*, *head grammars*, *indexed grammars*, *categorial grammars*, *simple matrix grammars*, etc. are well-known mechanisms that generate mildly context-sensitive languages.

Moreover, it is suggested natural languages could occupy an orthogonal position in the Chomsky Hierarchy (i.e., class of languages that contains some regular languages, some non-context-free, and so on). In fact, we can find some examples of natural languages constructions that are neither regular or context-free, and also some regular or context-free constructions that do not appear naturally in sentences. Thus, it seems that Chomsky Hierarchy is not the appropriate place for locating natural languages.

Therefore, if GI models aims to simulate the learning process of natural language they cannot focus on the inference of context-free or regular grammars, but it could be desirable that they concentrate on the learning of grammars that generates mildly context-sensitive languages and that occupies an orthogonal position in the Chomsky Hierarchy. In [11], it has been studied a non-classical mechanism with such properties: the class of *Simple p-dimensional External Contextual* grammars (SEC). Unlike the Chomsky grammars, SEC do not involve nonterminals and they do not have rules of derivation except one general rule: to adjoin contexts. Roughly speaking, a SEC produces a language starting from a word (*base*) and iteratively adding *contexts* (pair of words) at the ends of the currently generated word.

Becerra-Bonache and Yokomori [12] made the first attempt to learn SEC. They proved that the class of languages generated by SEC with fixed dimension p and fixed number of contexts q is learnable from positive data, from Shinohara's results. The learning algorithm derived from their main result was not time efficient. In [13], it was presented a polynomial-time algorithm for inferring SEC from positive data (small values of p and q were considered). Later, in [14], it was investigated for which choice of the parameter q (denoting the number of contexts) the class of SEC is iteratively learnable.

All these results suggest that SEC is an interesting class to study. Therefore, due to its linguistic and computational properties, SEC may be an appropriate candidate to model natural language syntax and could improve results in the field of GI.

5 Available Data for Learning

Another interesting question is to determine the data that must be available to the machine in order to learn the language. In order to correctly simulate natural language learning and take advantage of the simulation, the examples provided to our learning algorithm should be the same as the ones available to a child. But, which source of data is available to children during the learning process? This question has been a subject of

controversy and it is still of importance in discussions of learnability. Most GI systems are based on only positive data. If we look at natural language acquisition, the availability of positive data to children is trivially accepted. However, is this the only source of data available to children in order to acquire their native language?

Researchers tend to reduce the kind of data available to children to two types: positive and negative. Positive data is defined as sentences that are grammatically correct and all the remainder is considered negative data. The availability of positive evidence is widely accepted (children are exposed to a large amount of grammatical sentences uttered by adults), but the availability of negative evidence remains a matter of substantial controversy. The distinction between positive and negative data was used by Gold in [3]. This distinction is clear within the framework of formal languages, since positive data refers to strings that belong to the language and negative data to strings that do not belong. However, this classification seems not to be right within the framework of natural languages; it is difficult to classify all the data that children receive as positive or negative, since we can find sentences that are grammatically correct but contain negative information. Therefore, definitions about the kind of data available to children should be refined. Negative data should specially be well defined (its definition is so general than different interpretations have been given [15, 16]). Beliefs about whether or not children receive negative evidence depends crucially on how one defines that concept. Hence, it is important to define what negative data is exactly. If we consider that negative evidence is completely incorrect utterances from the adult, or adult replies to a child's ungrammatical utterance like "That's wrong", we can state that this source of evidence is very rare.

However, there is growing evidence that corrective input for grammatical errors is widely available to children ([17, 18]). During the first stages of children's language acquisition, adults tend to correct incomplete sentences uttered by the child; adults try to repeat the same idea but constructing grammatically correct sentences in the adult grammar [19]. This type of correction is called *expansion*. Expansion preserves the meaning of the child's utterance. Hence, adult's correction have the same meaning as the child's utterance, but different form. Moreover, the correction is a sentence that is grammatically correct, so positive information is obtained. Nevertheless, if a correction is received, this means that the string uttered by the child was not grammatically correct, so negative information is also obtained. Therefore, should corrections be considered as positive or negative data?

Most researchers have traditionally considered that corrections are negative data and should not be taken into account in the learning process, but as we can see, expansions are a kind of correction that is available to children (specifically during the two-word stage of child linguistic development, in which children go from the production of one word to the combination of two elements). Moreover, corrections are difficult to classify as positive or negative data, since they contain positive and negative information at the same time.

Although positive examples are an essential part of the language learning process and play the main role in that process, corrections can play a complementary role, providing additional information that can be helpful during the learning process. Moreover, the information available with a correction could improve learnability, and even some

aspects of the language could be learned faster. In fact, some studies show that children that receive expansions learn faster some aspects of the language than those that do not receive them [20].

Taking into account that none of the GI models has considered the combination of positive data and corrections, what we propose is to formalize a new learning model based on the combination of positive data and corrections. Moreover, this model will try to simulate the different stages of language acquisition in children, and to reflect the real interaction between child-adult during the process of language acquisition.

The first attempt to learn from corrections was made by Becerra-Bonache et al. in [21]. They applied the idea of corrections (given to the child during the process of language acquisition) to GI studies, and showed that models of GI can benefit from corrections, for instance, the query learning model proposed by Angluin [4]. Taking into account that the queries available to the learner in Angluin's model are quite unnatural for real learning environments, Becerra-Bonache et al. proposed a new type of queries called *correction query* (CQ). A CQ is defined as an extension of a MQ, but instead of a yes/no answer, a corrected string is returned to the learner; the correction consists of the shortest extension of the queried string. They proved that it is possible to learn DFA from corrections with a considerable reduced number of queries. Some other works have also followed this line of research, for example, [22]; they also use a correction based on the shortest extension of the wrong queried string, and showed the learnability of k-reversible regular languages in the limit. This model have been also applied to learning pattern languages.

In [23], a new CQ based on edit distance was introduced; when a string is submitted to the teacher, either he validates it (if it belongs to the target language), or he proposes a correction, that is to say, a string of the language close to the query with respect to the edit distance. In that way, the learner is corrected in a more "natural" way. Becerra-Bonache et al. proposed to learn classes of languages defined via edit distance (i.e., topological balls of strings), and with the help of this new CQ. They showed that this class is not learnable in Angluin's MAT model, but is with a linear number of CQs. Moreover, they conducted several experiments with a teacher simulating a human Expert, and showed that their algorithm is resistant to approximate answers. In [24], it is considered learning the class of pattern languages and a class of regular expressions using MQs and CQs also based in edit distance.

6 Syntax or Semantics?

Most of the research within the field of GI has focused on learning syntax, and tends to omit any semantic information. However, do children learn their native language independent of meaning? What is the role of semantics in language learning?

As linguistic and cognitive studies suggest, semantic and pragmatic information is also available to the child. Moreover, semantics and context seem to play an especially important role in the 2-word stage of child linguistic development. In this stage, context is important to understand the meaning of 2-word sentences and, thanks to the shared context, child and adult can communicate with each other although their grammars are different.

Taking into account that formal language learning is a hard problem, and taking into account the evidence of semantic learning in the first stages of natural language acquisition, we claim that semantic information can simplify the learning problem, and can make learning easier.

The first attempt to incorporate semantics in the field of GI has been made by Angluin and Becerra-Bonache in [25–27]. Inspired by the two-word stage of children's language acquisition, they proposed a computational model that takes into account semantics for language learning. In contrast to other approaches, their model does not rely on a complex syntactic mechanism; in that way, they try to represent the fact that, although the child and adult grammars are different, the semantic situation allows communication. This model also tries to give an account of the *meaning-preserving* corrections given to the child during the first stages of language acquisition (child's erroneous utterances are corrected by her parents based on the meaning that the child intends to express). This model has allowed them to investigate aspects of the roles of semantics and corrections in the process of learning to understand and speak a natural language.

7 Conclusions

How children acquire and use natural language is a fundamental problem that has attracted the attention of researchers for several decades. Besides obtaining a better understanding of natural language acquisition, interest in studying formal models of language learning stems also from the numerous practical applications of language learning by machines. In this paper we have proposed some ideas for simulating, in GI, the process of natural language acquisition. We claim that the simulation of that acquisition process might improve the methods in GI and, therefore, provide natural interfaces that may improve the efficiency and complexity of the mechanisms that we use in our everyday activities related with the information and the communication.

We have presented some ideas to improve models/techniques in GI by using as a model natural language acquisition. In that way, we have proposed that GI models use information that is relevant for natural language acquisition, that they take into account more aspects of real learning processes and use more natural tools. Thanks to these ideas, new challenging results in the field of GI can also be obtained. We have presented some works done in that direction; such works (bio-linguistically motivated) show that ideas coming from natural language acquisition studies can really improve results in the field of GI.

Therefore, ideas coming from linguistics can be useful in GI in order to obtain new perspectives of the problem and possible new solutions. But, of course, the theory of inferring formal grammars can also help to understand the process of language acquisition. GI can be relevant to understand language learning and could be a useful tool for any researcher interested in human language. Hence, the study of language learning from an interdisciplinary point of view is of great interest, not only to understand the learning mechanisms that underlie children's language acquisition, but also to develop computer systems that can recognize, understand and generate natural languages. In that way, such systems could also solve the problem of communication between machines and humans.

References

1. Parekh, R., Honavar, V.: Grammar inference, automata induction and language acquisition. In Dale, M., Somers, eds.: Handbook of Natural Language Processing. Marcel Dekker, New York, NY (2000) 727–774
2. Clark, A.: Grammatical inference and first language acquisition. In: Psychocomputational Models of Human Language Acquisition, Geneva (2004) 25–32
3. Gold, E.: Language identification in the limit. Information and Control 10 (1967) 447–474.
4. Angluin, D.: Learning regular sets from queries and counterexamples. Information and Computation 75 (1987) 87–106
5. Valiant, L.: A theory of the learnable. Communication of the ACM 27 (1984) 1134–1142
6. Bresnan, J., Kaplan, R., Peters, S., Zaenen, A.: Cross-serial dependencies in dutch. In Savitch, W., Bach, E., Marsh, W., Safran-Naveh, G., eds.: The Formal Complexity of Natural Language. D. Reidel, Dordrecht (1987) 286–319
7. Culy, C.: The complexity of the vocabulary of bambara. In Savitch, W., Bach, E., Marsh, W., Safran-Naveh, G., eds.: The Formal Complexity of Natural Language. (1987) 349–357
8. Pullum, G.K., Gazdar, G.: Natural languages and context-free languages. Linguistics and Philosophy 4 (1982) 471–504
9. Dassow, J., Păun, G.: Regulated Rewriting in Formal Language Theory. Springer-Verlag, Berlin (1989)
10. 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. Cambridge University Press, New York, NY (1985) 206–250
11. Becerra-Bonache, L.: On the Learnability of Mildly Context-Sensitive Languages using Positive Data and Correction Queries. PhD thesis, Rovira i Virgili University (2006)
12. Becerra-Bonache, L., Yokomori, T.: Learning mild context-sensitiveness: Toward understanding children's language learning. In Paliouras, G., Sakakibara, Y., eds.: ICGI. Volume 3264. Springer-Verlag, Berlin (2004) 53–64
13. Oates, T., Armstrong, T., Becerra-Bonache, L., 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. Volume 4201. Springer-Verlag, Berlin (2006) 137–147
14. Becerra-Bonache, L., Case, J., Jain, S., Stephan, F.: Iterative learning of simple external contextual languages. In: Proc. of ALT'08. Springer-Verlag, Berlin (2008) 359–373
15. Marcus, G.: Negative evidence in language acquisition. Cognition 46 (1993) 53–95
16. Saxton, M.: The contrast theory of negative input. Journal of Child Language 24 (1997) 139–161
17. Farrar, M.: Negative evidence and grammatical morpheme acquisition. Developmental Psychology 28 (1992) 90–98
18. Morgan, J., Bonamo, K., Travis, L.: Negative evidence on negative evidence. Developmental Psychology 31 (1995) 180–197
19. Chouinard, M., Clark, E.: Adult reformulations of child errors as negative evidence. Journal of Child Language 30 (2003) 637–669
20. Hernández Pina, F.: Teorías Psicosociolingüísticas y su aplicación a la adquisición del español como lengua materna. Siglo XXI, Madrid (1984)
21. Becerra-Bonache, L., Dediu, A.H., Tîrnauca, C.: Learning dfa from correction and equivalence queries. In Sakakibara, Y., Kobayashi, S., Sato, K., Nishino, T., Tomita, E., eds.: ICGI. Volume 4201. Springer-Verlag, Berlin (2006) 281–292
22. Tîrnauca, C., Knuutila, T.: Polynomial time algorithms for learning k-reversible languages and pattern languages with correction queries. In: ALT. Springer-Verlag, Berlin (2007) 264–276

23. Becerra-Bonache, L., de la Higuera, C., Janodet, J., Tantini, F.: Learning balls of strings from edit corrections. JMLR 9 (2008) 1841–1870
24. Kinber, E.: On learning regular expressions and patterns via membership and correction queries. In: ICGI. Springer-Verlag, Berlin (2008) 125–138
25. Angluin, D., Becerra Bonache, L.: Learning meaning before syntax. In: ICGI. Springer-Verlag, Berlin (2008) 1–14
26. Angluin, D., Becerra Bonache, L.: Experiments with an algorithm to learn meaning before syntax. In: ForLing2008. (2008) 1–12
27. Angluin, D., Becerra-Bonache, L.: A model of semantics and corrections in language learning. YALEU/DCS/TR-1425 (April, 2010)