Towards a Bio-computational Model of Natural Language Learning

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. This paper tries to bring together the theory of Grammatical Inference and the studies of natural language acquisition. We discuss how the studies of natural language acquisition can improve results in the field of Grammatical Inference, and how a computational model inspired by such studies can help to answer several key questions in natural language learning.

1 Introduction

Children, independently of their culture and the language they are exposed to, are able to acquire their native language easily, efficiently and without any specific training. However, the ease with which children acquire their language skills contrasts with the difficulty to explain this process.

The desire to better understand how children acquire their native language has motivated research in formal models of language learning [14,13]. Such models can allow us to address several key questions in natural language learning, such as: what types of input are available to the learner? what is the impact of semantic information on learning the syntax of a language? Moreover, such models can have important implications in the field of human language technologies. Therefore, it is of great interest to study formal models of language learning. However, in order to better simulate the human processing and acquisition of language, it is important that such models are inspired by studies of natural language acquisition.

Based on all these ideas, we present and discuss some of our main contributions within the field of *Grammatical Inference* (GI), and its implications for studies of natural language acquisition. We claim that ideas coming from studies of natural language acquisition can help GI to improve models and techniques used in this field, and even to obtain new challenging results. Moreover, GI models can also help in the understanding of natural language acquisition/processing. The simulation of such human capacity would provide natural interfaces to improve the communication between machines and humans.

J. Cabestany, I. Rojas, and G. Joya (Eds.): IWANN 2011, Part I, LNCS 6691, pp. 473–480, 2011.

2 Grammatical Inference versus Natural Language Acquisition

GI is a subfield of Machine Learning that deals with the learning of formal languages. 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 [10]. For example, imagine that the target language is ba^+ (i.e., a language that contains strings starting with one b, followed by at least one a), and the teacher provides to the learner strings that belong to the language (i.e., positive data), such as ba, baa, $baaa$... The learner, from this information, should infer that the target language is ba^+.

Several formal models of language learning have been proposed in the field of GI. The main ones are: *Identification in the limit* [11], *Query learning model* [1], and *PAC learning model* [15]. The problem of these models is that they do not take into account some relevant aspects of natural language learning. Therefore, they 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, let us focus on the model proposed by E.M. Gold, *Identification in the limit.* E.M. Gold was really motivated by the question of how children acquire their native language. His goal was to construct a formal model of language learning in order to investigate from a theoretical point of view how to learn a language artificially. *Identification in the limit* views learning as an infinite process. In this model, the learner passively receives more and more examples, and he has to produce a hypothesis of the target language (i.e., the language to be learned). If he receives new examples that are not consistent with the hypothesis, then he has to change it. The hypothesis has to converge to a correct hypothesis.

Note that in this model, the learner can never be sure of having correctly guessed the language, since new examples could appear at any time. The justification of Gold for studying this model was that

> *a person does not know when he is speaking a language correctly; there is always the possibility that he will find that his grammar contains an error. But we can guarantee that a child will eventually learn a natural language, even if it will not know when it is correct* [11]

Two traditional settings are considered: learning from *text* and learning from *informant.* In learning from text, only positive data (i.e., strings that belong to the language) are given to the learner. In learning from informant, positive and negative data are given to the learner (i.e., strings that belong and do not belong to the language).

Although we can find some similarities between learning in Gold's model and natural language acquisition (for example, in both cases there is a process of improvement), this model has several aspects that are controversial from a linguistic point of view.

In Gold's model, there is not limit on how long it can take the learner to guess the correct language.

> *That is, a language has been correctly identified when the learner no longer changes its guess through the presentation of all of the (possible infinite) strings in the language. If the learner is lucky, the first guess could be correct. Alternatively, it might take several billions of years to come up with the correct guess.*[12]

Hence, considerations of efficiency form a somewhat separate line of analysis from Gold's work. However, from natural language acquisition point of view efficiency is also important. Although learning natural language is an infinite process, we are able to learn the language in an efficient way.

Moreover, in *identification in the limit* the learner hypothesizes complete grammars instantaneously. From a linguistic point of view this assumption is unrealistic, since this is not the case in children's language acquisition.

Although natural language learning is mainly based on positive examples, positive data only is less than what a child actually gets in the learning process and, informant is much more than what a learner can expect. Moreover, the distinction that Gold does between positive and negative data is clear within the framework of formal languages, but not within the framework of natural languages, since we can find data that contains positive and negative information at the same time (as we will see in the next section), and hence it is difficult to classify them as positive or negative.

Learning from text or informant is also known as passive learning, as the learner passively received strings of the language. We know that natural language learning is more than that. Children also interact with their environment. They produce sentences that could be grammatically correct or not, and they can also ask questions to the adults, etc. Therefore, there is an interaction between child and adult, that is not gained by Gold's model.

Therefore, the model of *identification in the limit* postulates greatly idealized conditions, as compared to the conditions under which children learn language. For a longer discussion of the main models proposed in GI, see [6].

3 Towards a Bio-computational Model of Language Learning

The problem of language learning in GI presents similarities with the process of language acquisition. For example, in the context of natural language acquisition, 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 Chinese environment will learn to speak Chinese). Therefore, GI provides a good theoretical framework for investigating the process of language learning.

However, as we have seen in the previous section, the formal models proposed so far within the field of GI are not satisfactory. Therefore, it would be of great interest to develop a new computational model inspired by studies of natural language acquisition. We claim that studies of natural language acquisition can help to improve models and techniques in GI, and even to improve formal results obtained in this field. Moreover, a model with such bio-inspiration can also help to studies of natural language acquisition to better understand how children acquire and process their native language, and to answer several key questions about natural language learning (for example, what kind of data is available to the learner? what is the role of semantics in language learning?) The application of such a model could also be of great interest, for example, to improve the communication between humans and machines.

Next we present some works that we have done in these directions, and discuss the results obtained.

3.1 Learning from Corrections

Computational models of language learning should ideally provide the learner the same kind of information that are available to children. But, what kind of data is available to children during the learning process?

It is generally accepted that positive data are available to children. However, the availability of another kind of data, which is often called negative data, remains a matter of substantial controversy. As we can see, the definition of negative data is oversimplified, and could have different interpretations. Therefore, beliefs about whether or not children receive negative data are going to depend on how we define that concept. Moreover, as we have pointed in the previous section, this distinction between positive and negative seems not to be very adequate within the framework of natural languages, since we can find sentences that are grammatically correct but contain negative information.

For example, let us consider the following conversation extracted from the CHILDES database:

CHILD: milk, milk
ADULT: you want milk?
CHILD: uh-huh
ADULT: Ok. Just a second and I'll get you some

As we can see, the child produces a sentence that is grammatically incorrect. Immediately after, the adult tries to reformulate the sentence by checking on what the child had intended to say. Moreover, after that, the child acknowledges the reformulation.

This kind of conversations occurs very often during the first stages of children's language acquisition. Adults try to correct child's erroneous utterances based on the meaning that the child intend to express (then, the context in which this sentence is produced is very important). Adults correct them just to be sure that they have understood the child's intentions. Therefore, child's utterance and adult's correction have the same meaning, but different form.

What kind of data are these corrections? Positive or negative? As we can see, corrections contain positive and negative information at the same time. A correction is a sentence grammatically correct, then, contains positive information. But, as Chouinard and Clark pointed out

> *Since, like adults, children attend to contrast in form, any change in form that does not mark a distinct, different, meaning will signal to children that they may have produced something that is not acceptable in the target language.* [9]

Therefore, negative information is also available.

Based on linguistic arguments that support the presence of corrections in children's language acquisition [9], we have applied the idea of corrections to GI studies, and showed that GI models can be benefit from corrections, for instance, the query learning model proposed by D. Angluin. In this model, the learner is allowed to make queries to the teacher, and the teacher has to answer correctly his queries. Membership and equivalence queries have established themselves as the standard combination to be used. In the case of a membership query, the learner asks to the teacher if a string is in the target language, and the teacher answers "yes" or "not". In the case of a equivalence query, the learner asks if his conjecture is correct, and the teacher answers "yes" or gives to the learner a counterexample (if the conjecture is not correct).

The queries available to the learner in Angluin's model are quite unnatural for real learning environments. Based on the corrections that children receive during the first stages of language acquisition, we have proposed a new type of query called *correction query* (CQ). In a CQ, the learner asks if a string is in the language, and if the string does not belong to the language, the teacher returns a correction.

In [8], we present the first attempt to learn from corrections. Taking into account the simplicity of DFA and their adequacy for some applications of natural language processing, we considered that a good starting point was to apply corrections to learn deterministic finite automata (DFA). We design an algorithm called *Learning from Corrections Algorithm* (LCA), which is able to infer a DFA using CQs and equivalence queries. In this context, a correction of a string consists of the shortest extension of the queried string. We showed that it is possible to learn DFA from corrections, and that the number of queries needed by the learner is reduced considerably.

In [7] we proposed a new CQ based on edit distance. When the learner submits to the teacher a string that does not belong to the target language, the teacher returns a string of the language close to the query with respect to the edit distance (the edit distance is the minimum number of deletion, insertion or substitution operations needed to transform one string into another). We consider non-standard classes of languages defined via edit distance : the balls of strings. We showed that this class is not learnable in Angluin's model, but is with a linear number of CQs. We also conducted several experiments with a teacher simulating a human Expert, and showed that our algorithm is resistant to approximate answers.

Therefore, all these results show that new challenging results can be obtained in the field of GI by using ideas coming from studies of natural language acquisition.

3.2 Learning with Semantics

The kinds of corrections considered in the papers cited above, are mainly syntactic corrections based on proximity between strings. However, as we have pointed out in the previous section, the corrections given to children during the first stages of language acquisition *preserve* the meaning that the child intend to express. Therefore, one of our goals has been to develop a formal model that gives an account of this kind of correction, 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?

Inspired by the two-word stage of children's language acquisition, we have developed a formal model that takes into account semantics for language learning. This model accommodates two different tasks: comprehension and production. Such a model tries to reflect several aspects of natural language acquisition. For example, our model does not rely on a complex syntactic mechanism; in that way, we try to represent the fact that, although the child and adult grammars are different, the semantic situation allows communication

The first attempt to incorporate semantics in the field of GI can be found in [3,2,5]. We have presented an algorithm that learns a meaning function and prove that it finitely converges to a correct result under a specific set of assumptions about the transducer and examples used. The learning problem has been formulated as follows: (i) The teacher provides to the learner several example pairs consisting of a situation and an utterance denoting something in the situation (like in the 2-words stage, where in addition to hearing utterances, children have access to the context in which these utterances are generated); (ii) The goal of the learner is to learn the meaning function, allowing the learner to comprehend novel utterances. We have shown that a simple algorithm can learn to comprehend an adults utterance (in the sense of producing the same sequence of predicates), even without mastering the adults grammar. We have presented and analyze the results of empirical tests of our algorithm with natural language samples in an example domain of geometric shapes and their properties and relations.

We have also explored the possibility of applying existing automata-theoretic approaches to machine translation (concretely, subsequential transducers and the OSTIA algorithm) to model language production [4]. For ten natural languages and a limited domain of geometric shapes and their properties and relations we have defined sequential transducers to produce pairs consisting of an utterance in that language and its meaning. Using this data we have empirically explored the properties of OSTIA and DD-OSTIA algorithms for the tasks of learning comprehension and production in this domain, to assess whether they may provide a basis for a model of meaning-preserving corrections. Our results suggest that OSTIA and DD-OSTIA may be an effective method to learn to translate sequences of predicates into natural language utterances in our domain. However, some of our objectives seem incompatible with the properties of OSTIA (e.g., the intermediate results of the learning process do not seem to have the properties we expect of a learner who is progressing towards mastery of production).

Finally, we have considered a statistical approach to model comprehension and production, which has produced a more powerful version of our initial model and has allowed us to model corrections [5]. In this new approach, the teacher is able to understand a flawed utterance produced by the learner and respond with a correct utterance for that meaning. Moreover, the learner can recognize that the teachers utterance has the same meaning but different form. This approach allows us to compare a learner that only receives positive data, a learner that is corrected sometimes (with different probabilities) and a learner that is corrected whenever this is possible, and therefore, to study the effect of meaning-preserving corrections on language learning. The results obtained so far show that: the access to the semantics facilitates language learning, and the presence of corrections by the teacher has an effect on language learning by the learner (even if the learner does not treat corrections specially). Hence, this new approach points out the relevance of semantics and corrections in language learning, and sheds interesting questions about them.

4 Conclusions

The understanding and simulation of natural language acquisition constitutes one of the biggest challenges of the 21st century. Therefore, it is of great interest to develop formal models of language learning that can help us to better understand how children acquire their native language. Such a models could also have important implications in the field of human language technologies. If we are able to create machines that can recognize, understand and generate natural languages, 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.

In this paper we have discussed how the theory of GI and the studies of natural language acquisition can be brought together. Based on the fact that language learning in GI exhibits similarities with natural language acquisition, the need for an adequate/sophisticated bio-computational model for language learning has been discussed and confirmed.

To employ the results of the field of GI and natural language acquisition in each other, both theories should be developed. On one hand, we have argued why GI models need to involve the concept of corrections, and we have demonstrated how the models get advantage in this way. On the other hand, we have shown how a computational model that incorporates semantics as well, can allowed us to investigate aspects of the roles of semantics and corrections in the process of learning to understand and speak a natural language.

Therefore, we have tried to bring together the theory of GI and studies of natural language acquisition, and shown the benefits that can be obtained by doing so. Ideas coming from linguistics can be useful in GI in order to obtain new perspectives of the problem and possible new solutions and, of course, the theory of GI can also help to understand the process of language acquisition. Hence, it is of great interest to study natural language acquisition from an interdisciplinary point of view. Ideas and techniques coming from different areas can help us to develop computer systems with human-like capabilities and go deeper in the understanding of children's language acquisition.

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. 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: ICGI, pp. 1–14. Springer, Berlin (2008)
5. Angluin, D., Becerra-Bonache, L.: Experiments using OSTIA for a language production task. In: CLAGI 2009, pp. 16–23 (2009)
6. Becerra-Bonache, L.: On the Learnability of Mildly Context-Sensitive Languages using Positive Data and Correction Queries. PhD thesis, Rovira i Virgili University (2006)
7. Becerra-Bonache, L., de la Higuera, C., Janodet, J.C., Tantini, F.: Learning balls of strings from edit corrections. JMLR 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. Chouinard, M.M., Clark, E.V.: Adult reformulations of child errors as negative evidence. Journal of Child Language 30, 637–669 (2003)
10. Clark, A.: Grammatical inference and first language acquisition. In: Psychocomputational Models of Human Language Acquisition, Geneva, pp. 25–32 (2004)
11. Gold, E.M.: Language identification in the limit. Information and Control 10, 447–474 (1967)
12. Gordon, P.: Learnability and feedback. Developmental Psychology 26, 217–220 (1990)
13. Parekh, R., Honavar, V.: Grammar inference, automata induction and language acquisition. In: Moisl, Dale, S. (eds.) Handbook of Natural Language Processing, pp. 727–774. Marcel Dekker, New York (2000)
14. Pinker, S.: Formal models of language learning. Cognition 7, 217–283 (1979)
15. Valiant, L.G.: A theory of the learnable. Communication of the ACM 27, 1134–1142 (1984)