Learning, Agents, and Formal Languages: Linguistic Applications of Interdisciplinary Fields

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

[1] Laboratoire Hubert Curien, Jean Monnet University,
18 rue Benoit Lauras, 42100, Saint-Etienne, France
leonor.becerra@univ-st-etienne.fr
[2] Research Group on Mathematical Linguistics,
Universitat Rovira i Virgili, Av. Catalunya 35, 43002 Tarragona, Spain
mariadolores.jimenez@urv.cat

Abstract. This paper focuses on three areas: machine learning, agent technologies and formal language theory. Our goal is to show how the interrelation among agents, learning and formal languages can contribute to the solution of a challenging problem: the explanation of how natural language is acquired and processed. Linguistic contributions of the intersection between machine learning and formal language theory –through the field of grammatical inference– are reviewed. Agent-based formal language models as colonies, grammar systems and eco-grammar systems have been applied to different natural language issues. We review the most relevant applications of these models.

1 Introduction

Nowadays, interdisciplinary research is key to make progress and increase the rate of scientific findings in different areas. There are problems that cannot be approached just by the single perspective of a specific field. Therefore, to understand better or solve these kind of problems, the collaboration of researchers from different disciplines is required.

Taking into account the relevance of interdisciplinarity, we consider the relationship among *machine learning, formal languages* and *agent technologies*:

- **Machine Learning** is one of the most active research areas within Artificial Intelligence. Its main goal is to develop techniques that allow computers to learn. Machine learning algorithms construct a model based on the inputs that they receive, and then they use that model to make predictions or decisions. Examples applications of machine learning are: spam filtering, handwriting recognition, computer vision, etc.
- **Formal Languages** was originated from mathematics (researchers as Thue, Post, Turing) and linguistics (Chomsky). Formal language theory provides mathematical tools for the description of linguistic phenomena. It was born in the middle of the 20th century as a tool for modeling and investigating the syntax of natural languages. However, very soon it developed as a new research field, separated from Linguistics, with specific problems, techniques and results and, since then, it has had an important role in the field of Computer Science.

J. Bajo et al. (eds.), *Trends in Prac. Appl. of Agents, Multi-Agent Sys. and Sustainability*,
Advances in Intelligent Systems and Computing 372, DOI: 10.1007/978-3-319-19629-9_5

- **Agent Technologies** is one of the most important areas emerged in Information Technology in the 90's. By implementing autonomous entities driven by beliefs, goals, capabilities, plans and agency properties, agent technologies capture essencial aspects of the modeled systems. The metaphor of autonomous problem solving entities cooperating and coordinating to achieve their objectives is a natural way of conceptualizing many problems. In fact, the multi-agent system literature spans a wide range of fields.

Our main goal here is to review the *linguistic* contributions of the interrelation of those three areas. This is, we want to show how the interdisciplinary relation between machine learning, formal languages and agent technologies can contribute to the solution of one of the most persistent problems in science: the explanation of how natural language is acquired and processed.

If we want to explain natural language, we need to cross traditional academic boundaries in order to solve the different problems related to this topic. We should attack the subject from various angles and methods, eventually across disciplines, forming a new method for understanding natural language. Therefore, interdisciplinarity should be an essential trait of the research in language. In this paper, we review some interdisciplinary research performed in this direction. Specifically, in section 2, the intersection between machine learning and formal languages is taken into account. This intersection constitutes a well-established research field known as *Grammatical Inference*. We review here the main linguistic applications of this field. In section 3, the interrelation between formal language theory and agent technologies is considered. We show how agent technologies can offer good solutions and alternative frameworks to classic models in the area of processing and computing languages that can be useful for the description and analysis of natural language.

2 Learning and Formal Languages

The intersection between the field of *machine learning* and the *theory of formal languages* constitutes a well-known research field called *Grammatical Inference* (GI) [13]. GI studies how grammars can be learnt from a set of data. The field of GI was originated in the 60's, mainly by the work developed by E.M. Gold [15]. Motivated by the problem of how children acquire their native language, E.M. Gold tried to investigate, from a theoretical point of view, how the ability to speak a language can be achieved in an artificial way. Since then, a big amount of research has been done by researchers coming from different scientific traditions: machine learning, formal languages, computational linguistics, pattern recognition, etc. Two main approaches can be distinguished in GI: 1) *Theoretical approaches*: researchers aim to prove efficient learnability of grammars. Most of GI researchers have been focused on this approach. Their aim is to obtain formal results, for example: formal descriptions of the target languages, formal proofs about the efficiency of a learning algorithm, etc.; and 2) *Practical approaches*: researchers aim to develop systems that learn grammars from real data. Instead of proving the learnability of grammars, researchers focus on providing empirical systems that deal with natural language data.

Despite the original linguistic motivation of the GI studies, most of the work in this field have been focused on obtaining formal results, without exploiting the linguistic relevance of the classes that have been studied and the results obtained. We will review next some practical studies in GI based on natural language data.

2.1 Practical Studies with Natural Language Data

Although most part of the work in GI are theoretical, we can also find some practical approaches in GI based on natural language data. Next we review some of the main approaches.

First of all, it is important to point out that in a GI problem, there are two different actors involved: a teacher and a learner. The teacher provides information to the learner, and the learner (or learning algorithm), from that information, must identify the underlying language. Depending on how this information is provided to the learner, we can distinguish three different GI approaches to natural language [14]: 1) *Unsupervised approach*: the teacher provides *unlabeled* examples to the learner, that is, the learner does not receive explicit information about the structure of the sentences in the target language.; *Supervised approach*: the input data consists of a set of *labeled* examples, that is, the learner receives examples of inputs paired with the corresponding correct outputs; and 3) *Semi-supervised approach*: in addition to the *labeled* examples, the learner receives *unlabeled* examples.

Like theoretical studies developed in GI, most part of GI methods for natural language have also been focused on CF grammars. Most of these natural language learning methods are based on an *unsupervised approach*, and only use *positive data* during the learning process (i.e., the learner only receives examples of sentences that belong to the target language). The most common method to evaluate these systems is by using a treebank (i.e., a linguistic corpus in which sentences are annotated with their syntactic structure, often represented in the form of a tree). This method consists in extracting from a treebank (selected as the "gold standard") a set of plain natural language sentences and giving it to the algorithm as input. Then, the GI algorithm generates structured sentences and these sentences are compared against the original structured sentences from the treebank. Different metrics can be used to do this comparison, but the most used are precision (which shows how many learned structures are correct, describing in that way the correctness of the learned grammar) and recall (which shows how many of the correct structures have been learned, giving in that way the completeness of the learned grammar). It is worth noting that one of the most used treebanks is ATIS (Air Traffic Information System), an English corpus that contains mostly questions and imperative sentences on air traffic. Examples of GI systems for natural language learning are: EMILE [1], ABL (Alignment-Based Learning algorithm) [20] and ADIOS (Automatic DIstillation Of Structure algorithm) [19]. For detailed information about these algorithms and other grammar inference methods, see [14].

In GI there has also been some efforts for taking into account more natural aspects during the language learning process. An example of it is [2,3,4]. The purpose of this work was not to learn a CF grammar, but to investigate the effect of semantics and corrections in the process of learning to understand and speak a natural language. In fact, in the early stages of children's linguistic development, semantic information seem to

play an important role [10]. Taking into account that most works in GI have only been focused on syntax learning and results in GI show that language learning is hard, the following question was posed: can semantic information simplify the learning problem? In order to answer this question, a simple computational model was developed which takes into account semantics for language learning. The model was tested with ten different natural languages by using a simplified version of the Miniature Language Acquisition task (this task involves sentences that describe visual scenes), and the results show that: i) access to semantic information facilitates language learning; ii) the presence of meaning-preserving corrections has an effect of language learning, even if the learner does not treat them specially. Therefore, the results obtained with this work were not only of interest for GI, but also for the linguistic community.

Researchers in GI have also developed methods for other tasks, such as machine translation. For example, we can find several works focused on learning stochastic finite-state transducers (SFST) for machine translation [9]. A SFST involves two different alphabets (source and target alphabet), and associates probabilities to the transitions and final states. The main advantages of working with these models are the following: i) there exist efficient search algorithms for translation; ii) these techniques are less computational expensive (in general) than most pure statistical approaches; iii) they allow an easy integration with other information sources, such as acoustics models, making easier the applications of SFST to more difficult tasks, such as speech translation. It is worth noting that these methods have been successfully applied to different non-trivial tasks, such as Miniature Language Acquisition task, Traveler task, etc. For more information, the reader is refered to [9].

3 Agents and Formal Languages

In this section, we consider the interrelation between *multi-agent systems* and *formal language theory*. The first generation of formal grammars, based in rewriting, formalized classical computing models. At that time, linguistics was the central application of formal language theory and linguists were very much interested in applying formal language models to the formalization of natural language. However, from the 90s, the interest of linguistics in formal languages seems to have disappeared and formal language theorists have found innumerable applications of their theory different from linguistics. Problems related to the first generation of formal languages based on rewriting systems were the reason for that divorce between formal languages and linguistics. However, models proposed from the 90s in the area of formal language theory may solve those classic problems. Among those new models we find the *agent-based models* of formal languages that constitute an important subfield of the theory. The main advantage of those agent-based models is that they increase the power of their component grammars thanks to interaction, distribution and cooperation. *Colonies*, *grammar systems* and *eco-grammar systems* are examples of this new generation of formal languages. All these new types of formalisms have been proposed as grammatical models of agent systems. These multi-agent formal languages have been applied to natural language description and processing. In general, it can be shown that those non-standard models in formal languages can solve the classical problems related to the first generation of formal languages and can cover the whole range of linguistic disciplines, from

phonology to pragmatics. In the next section we show some examples of these possible linguistic applications.

3.1 Linguistic Applications of Agent Based Formal Language Theory

Colonies are the first agent-based model we consider. Colonies as well-formalized language generating devices have been proposed in [17], and developed during the nineties in several directions. Colonies can be thought of as grammatical models of multi-agent systems motivated by Brooks' subsumption architectures [8]. They describe language classes in terms of behavior of collections of very simple, purely reactive, situated agents with emergent behavior. Colonies, as proposed originally, capture some formal aspects of systems of finite number of autonomous components capable to perform very simple reactive computing tasks each. The behaviour of the colony really emerges from interactions of its components with their symbolic environment and can considerably surpass the individual behaviours of its components. The main advantage of colonies is their generative power, the class of languages describable by colonies that make use of strictly regular components is beyond the set describable in terms of individual regular grammars.

In [7], colonies have been proposed as a tool to generate natural language by the interaction of a finite number of finite-state devices that generate finite languages. This application takes into account the idea of describing natural languages as a number of modules that interact in a nonsimple way. Colonies offer a modular theory where the various dimensions of linguistic representation may be arranged in a distributed framework and where the language of the system is the result of the interaction of those independent cooperative modules. Colonies allow us to generate infinite languages by only using grammars generating finite languages. This formal framework increases the power of regular grammars thanks to interaction. What is important here is the fact that although the generative power of colonies goes beyond the regular family of languages, the derivation process is done in a regular (finite-state) manner. Therefore, colonies may reveal as a device able of conjoining the simplicity of finite-state machines with a stronger generative power able to account for the infiniteness (context-free or more) of natural languages.

Another important agent based-model in formal languages are the so-called *grammar systems* [11]. Grammar system theory is a consolidated and active branch in the field of formal languages that provides syntactic models for describing multi-agent systems at the symbolic level, using tools from formal grammars and languages. A grammar system is a set of grammars working together, according to a specified protocol, to generate a language. Note that while in classical formal language theory *one* grammar (or automaton) works individually to generate (or recognize) *one* language, here we have *several* grammars working together in order to produce *one* language. The theory was launched in 1988 and has developed into several directions, motivated by several scientific areas.

Grammar systems may offer useful tools to account for arrangement and interaction of the various dimensions of natural language grammar. In order to define a grammar system approach for grammar architecture, a set of postulates of linguistic theory must be followed [18]: 1) we take a grammar to be a set of subgrammars called modules; 2)

each of these modules is a grammar of an independent level of linguistic representation; these modules are not hierarchically related to one another; 3) a module need not wait for the output of another to do its work, but has the power to generate (analyze) an infinite set of representations quite independently of what is going on in any of the other components; 4) each component is a self-contained system, with its own independent set of rules, principles and basic vocabulary; 5) the lexicon plays a special, transmodular role in the theory. In order to capture every feature of the above list, a new variant of grammar systems has been introduced: *Linguistic Grammar Systems* (LGS) [16]. LGS offer an example of the possible application of formal languages to linguistics. For formal definitions of LGS the reader can see [16].

Very relevant in the area of multi-agent models of formal languages are the so-called *eco-grammar systems* [12]. Eco-grammar systems provide a syntactical framework for eco-systems, this is, for communities of evolving agents and their interrelated environment. An eco-grammar system is defined as a multi-agent system where different components, apart from interacting among themselves, interact with a special component called 'environment'. Within an eco-grammar system we can distinguish two types of components *environment* and *agents*. Both are represented at any moment by a string of symbols that identifies the current state of the component. These strings change according to sets of evolution rules. Interaction among agents and environment is carried out through agents' actions performed on the environmental state by the application of some productions from the set of action rules of agents.

Eco-grammar systems can model the structure of dialogue and account for the evolution of language. According to the idea that dialogue can be understood as the sustained production of mutually-dependent acts, constructed by two or more agents each monitoring and building on the actions of the other, eco-grammar systems can describe dialogue as a sequence of *acts* performed by two or more agents in a common environment. An example of this application is presented in [6]. In this paper, a formal model of dialogue based on eco-grammar systems is introduced: *Conversational Grammar Systems* (CGS). CGS present some advantages to account for dialogue: a) the generation process is highly *modularized* by a distributed system of contributing agents; b) it is *contextualized*, linguistic agents re-define their capabilities for acting according to context conditions given by mappings; c) and *emergent*, it emerges from current competence of the collection of active agents.

4 Conclusion

We have considered three different areas: agent systems, machine learning and formal languages. Each of those fields has intrisically good features that enable them to cope with many real-world problems. The formal apparatus of agent technology provides a powerful and useful set of structures and processes for designing and building complex applications. Machine learning is one of the core fields of Artificial Intelligence since the ability to learn is one of the most fundamental attributes of intelligent behavior. And finally, formal language theory provides the flexibility and the abstraction necessary in order to be applied to fields such as linguistics, economic modeling, developmental biology, cryptography, sociology, etc. Therefore, multi-agent systems, machine learning

and formal language theory provide flexible and useful tools that can be used in different research areas due to their versatility. In this paper, we have tried to show that the individual power of those systems may be increased if they collaborate among them.

In this work, we have focused on two possible intersections: formal languages and learning; and, agents and formal languages. Our main objective here has been to show the contributions of those intersections to the area of natural language processing. The interaction between researchers in those three topics can provide good techniques and methods for improving our knowledge about how languages are processed.

Language is one of the most challenging issues that remain to be explained. Natural language is a hard problem not only for linguistics that has not yet provided universal accepted theories about how language is acquired and processed, but also for computer science that up to now has implemented natural language processing systems that are far from being satisfactory. As a complex system, the explanation, formal modelling and simulation of language present important difficulties. If we deal with language, we need to connect and integrate several academic disciplines in order to find a solution. In this interdisciplinary environment, formal languages, machine learning and agents systems can collaborate –as shown in this paper– in the description, explanation and processing of language.

References

1. Adriaans, P.: Language learning from a categorial perspective. PhD thesis, University of Amsterdam (1992)
2. 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)
3. Angluin, D., Becerra-Bonache, L.: Effects of meaning-preserving corrections on language learning. In: CoNLL 2011, pp. 97–105 (2011)
4. Angluin, D., Becerra-Bonache, L.: A Model of semantics and corrections in language learning. Technical Report, Yale University, 1–45 (2010)
5. Becerra-Bonache, L., Case, J., Jain, S., Stephan, F.: Iterative learning of simple external contextual languages. Theoretical Computer Science 411, 2741–2756 (2010)
6. Bel-Enguix, G., Jiménez-López, M.D.: Modelling dialogue as inter-action. International Journal of Speech Technology 11(3/4), 209–221 (2008)
7. Bel-Enguix, G., Jiménez-López, M.D., Martín-Vide, C.: Using finite-state methods for getting infinite languages: A preview. Romanian Journal of Information, Science and Technology 12(2), 125–137 (2009)
8. Brooks, R.A.: Elephants don't play chess. Robotics and Autonomous Systems 6, 3–15 (1990)
9. Casacuberta, F., Vidal, E.: Learning finite-state models for machine translation. Machine Learning 66(1), 69–91 (2007)
10. Chouinard, M.M., Clark, E.V.: Adult reformulations of child errors as negative evidence. Journal of Child Language 30, 637–669 (2003)
11. Csuhaj-Varjú, E., Dassow, J., Kelemen, J., Păun, G.: Grammar systems: A grammatical approach to distribution and cooperation. Gordon and Breach, London (1994)
12. Csuhaj-Varjú, E., Kelemen, J., Kelemenová, A., Păun, G.: Eco-grammar systems: A grammatical framework for life-like interactions. Artificial Life 3(1), 1–28 (1996)
13. de la Higuera, C.: Grammatical inference: Learning automata and grammars. Cambridge University Press, Cambridge (2010)

14. D'Ulizia, A., Ferri, F., Grifoni, P.: A survey of grammatical inference methods for natural language learning. Artificial Intelligence Review 36(1), 1–27 (2011)
15. Gold, E.M.: Language identification in the limit. Information and Control 10, 447–474 (1967)
16. Jiménez-López, M.D.: A grammar systems approach to natural language grammar. Linguistics and Philosophy 29, 419–454 (2006)
17. Kelemen, J., Kelemenová, A.: A grammar-theoretic treatment of multiagent systems. Cybernetics and Systems 23, 621–633 (1992)
18. Sadock, J.M.: Autolexical syntax. A theory of parallel grammatical representations. University of Chicago Press, Chicago (1991)
19. Solan, Z., Horn, D., Ruppin, E., Edelman, S.: Unsupervised learning of natural languages. PNAS 102(33), 11629–11634 (2005)
20. van Zaanen, M.: Bootstrapping structure into language: alignment-based learning. PhD thesis, University of Leeds (2001)