Emergent Generalization in Bayesian Agents Using Iterated Learning

Giancarlo Schrementi and Michael Gasser

School of Informatics and Computing
Indiana University, Bloomington, IN 47405
gischrem@cs.indiana.edu

Abstract

The compositional nature of human language is a remarkable adaptation that solves the problem of generalizing our communications to novel experiences. The Iterated Learning Model of agent interaction has proven to be a useful tool for exploring the emergence of this phenomenon of generalization. Recently, a Bayesian interpretation of this model has been proposed and analyzed in the literature. The work here combines the Bayesian approach with the traditional goal of iterated learning, the emergence of compositional communication. Two methods of measuring language likelihood are investigated, one based on agent comprehension and the other on production scope. Calculating likelihood based on agent comprehension is shown to result in the emergence of significantly better generalization. The beneficial effect of a description-length based prior probability is also demonstrated.

Introduction

The ability to generalize our knowledge to novel experiences is a fundamental capability of the human mind. Nowhere has this faculty had more impact than on how we communicate. Our languages have developed to be massively compositional. As children, we learn a set of components and rules for combining those components in a way that allows us to express an infinite number of utterances. Likewise, we can understand those utterances by breaking them down into their components and rules. Thus the compositional nature of our languages has given us tremendous ability to generalize our communications.

In this paper, we look at how this compositional nature emerges through communicative interactions between agents that are finite-state transducers. In order to model these interactions, we use the Iterated Learning Model (ILM) of Kirby and colleagues (Kirby, 2001; Smith et al., 2003). ILM originated as way to model this kind of language emergence and evolution, but has since been used as a more general model of knowledge change in domains with a teacher and a learner (Kalish et al., 2007).

Iterated learning can involve many agents, but in its purest form involves a single teacher and a single learner. Initially, the teacher agent imparts some of its knowledge to a learner. Since the teacher is not revealing all of its knowledge, the learner must fill in the blanks according to some inference algorithm. Typically, the inference algorithm looks at the knowledge the learner already has and infers from that. The learner then becomes a teacher and instructs a new learner in the same fashion and this continues for many iterations. Eventually this process of knowledge transfer and self-organization converges to an equilibrium in a manner similar to the transfer and self-organization of genetic information in an artificial life simulation.

Language evolution models usually operate with a space of idealized meanings that agents need to communicate to each other. These meanings take the form of vectors of features, each having some range of values. The agents then turn these meanings into some form of signal, creating a meaning-signal mapping. In iterated learning models, the agents can be broadly defined to fall into two categories.

The first type of agent we will call grammatical inducers. These grammatical inducers keep track of any correlations between features in the meaning space and the received signals. These correlations are kept track of with a context-free grammar, neural network, or matrix. The agents induce a signal for a novel meaning by making use of any noticed correlations between the features of the meaning and portions of earlier signals. Those correlations are typically combined with a randomly generated signal portion that represents the rest of the uncorrelated features to create a final signal for the novel meaning. The success of these agents is judged by how compositional their signals are after a number of generations. Originally this was measured through subjective analysis of the signals (Batali, 1998), but more recently is often measured by $expressivity$ (Kirby, 2007; Brighton, 2005). Expressivity is defined as the number of meanings that can be distinctly expressed.

The second type of agent is the more recent Bayesian agent that was analyzed in detail by Griffiths and Kalish (2005, 2007). Griffiths recognized that the learner in ILM is essentially using a form of Bayesian inference to infer the language from the teacher's instruction. The learner considers many hypotheses about the language before picking

the one that it feels is most probable. The probability of a hypothesis is calculated based on how closely a hypothesis matches the data from the teacher, the likelihood, and by the agent's inductive biases, the prior. This relationship allows iterated learning to be formulated as a mathematical process that can be rigorously analyzed. One of the results of Griffiths' analysis is that over generations of iterated learning the posterior probability distribution converges to the prior probability distribution. Essentially, the languages the inductive biases favor are the languages that will emerge over the course of the process.

The convergence of the Bayesian agent form of ILM has been rigorously analyzed (Rafferty et al., 2009; Ferdinand and Zuidema, 2009). However, these studies have used arbitrary priors and were not looking for evidence of compositionality in agent signals. The work here combines the goals of the grammatical inducers with the method of the Bayesian inducers. To do this we need to characterize what information our prior is to use and how to calculate likelihood.

Bayesian inference, Equation (1), has long been known to be related to the mathematical model selection criterion of Rissanen (1978) called the Minimum Description Length Principle (MDL) and the closely related Minimum Message Length (MML) measurement of Wallace and Boulton (1968). A detailed discussion of this relationship is in Vitanyi and Li (2000), but we will discuss the nature of the correspondence here.

$$P(Model|Data) = \frac{P(Data|Model)P(Model)}{P(Data)} \quad (1)$$

Both MDL and MML measure the success of a mathematical model of data. A successful model is one that is simple and compactly expresses the data. By combining a measure of the size of the model and a measure of the size of the data as encoded by the model the total information load can be quantified. The essence of the relationship with Bayesian inference is that the amount of information can be viewed as the amount of Shannon entropy. A higher information load corresponds to a model with lower posterior probability, $P(Model|Data)$. The relationship extends to the two primary components of Bayesian inference, the likelihood and the prior. The likelihood, $P(Data|Model)$, corresponds to the size of the data as encoded by the model and the prior, $P(Model)$, corresponds to the complexity of the model.

The selective pressures of minimizing description length on a model are not very different from the selective pressures on a language. Language is a model that uses syntax to represent semantics. A successful language is one that can express everything we want to talk about but is also simple to learn and use. This correspondence provides us with a way to formulate the Bayesian inference components of our agents. The likelihood needs to measure how successful we are at expressing ourselves and the prior needs to measure how simple our manner of expression is.

This is not the first time MDL is used as a way to encourage to the emergence of generalization without directly selecting for it. Schrementi and Gasser (2010) used it as a fitness metric for a genetic algorithm. Brighton (2005, 2003, 2002) used description length as a hypothesis selection measure in an iterated learning model that used a modified form of transducers called finite-state unification transducers. Brighton's work was not specifically Bayesian and stayed close to the original formulation of the likelihood in MDL; that likelihood was the size of the data as encoded by the model. The focus of the work here is to investigate likelihood as a measure of the probability that a signal can be decoded to its original meaning. We investigate two methods of formulating likelihood as a probability, one based on expressivity and the other comprehension.

Iterated Learning Framework

Our implementation of the iterated learning model uses agents that are simple finite-state transducers. These transducers sequentially process input strings and encode them into output strings. Each edge between states in the transducer reads in an input character and writes an output character. This encoding process provides a simple way to model linguistic production, the translation of meaning into signal.

Notably, the same transducer can be used for the other half of a linguistic interaction, comprehension, by reversing what is read and what is written for each edge. This inverted transducer will be able to translate the output strings back into the original input strings, with an important caveat. The inversion process can introduce ambiguity in the transducer that didn't exist before. A state that has two edges leaving it that output the same character will after inversion have two edges leaving it that read the same character. This ambiguity results in a non-deterministic transducer that can have multiple paths that read the same input string.

The algorithm starts with a state-minimal finite-state automaton that recognizes the entire set of input training strings. A transducer recognizes a string if it finishes in an end state after reading the string. A state-minimal transducer is one that has been compressed to have the fewest states needed to recognize the input set and only the input set. Each edge in the automaton is then randomly assigned to write one of the output characters. This transducer is the first teacher in the iterated learning process. The learner starts out as an empty transducer, with just a start state and an end state.

The learning process begins with the teacher going through a random selection of the input training strings and producing an input-output pair. The learner adds each of these input-output pairs to its transducer, such that there is a path from the start state to the end state that reads the input string and writes the output string. Any remaining input

training strings are added to the learner but not paired with any output. The learner's transducer is then compressed to be state-minimal. This results in a learner that has the same transducer structure as the teacher but some of the edges may not write anything.

The edges that have no output form the basis for the invention part of the iterated learning model. Invention refers to the process of inferring outputs for inputs which were not presented to the agents as input-output pairs by their teacher. Our invention method uses Bayesian inference to select the output characters for the edges that lack them. The set of sets of possible outputs to fill in the blanks forms a search space whose size is determined by the number of blanks, n, and the size of the output character space. For each of the experiments in this paper, there are two possible output characters, resulting in a space of 2^n.

Each set of output characters in the search space is a hypothesis of the optimal language. This hypothesis coupled with the learned transducer completely specifies all the input-output mappings of the agent for the training strings. The transducer can now be further compressed following a compression criterion from Brighton (2002). The criterion is that any two states can be combined if the change doesn't affect the input-output mappings of the training strings. We have added two additional criteria. The first is that the two states don't have conflicting output edges, e.g. two edges reading the same character but writing a different character, which prevents production ambiguity. The second is that the two states to be combined must also be at the same depth from the initial state, in order to prevent cycles and to allow the compression to be done iteratively.

The further compressed transducer now recognizes and encodes additional strings beyond those that it was trained on. In essence, this compression allow the transducer to generalize its knowledge about the training set to a wider range of input strings. Each hypothesis results in a transducer that can be compressed in this way to different degrees. The size of this compressed transducer will form the basis of our calculation of the prior probability of a hypothesis. Additionally, we can now measure how well a given language, as specified by the transducer, generalizes to novel test strings.

The posterior probability of each hypothesis in the search space is calculated according to our formulation of its prior probability and likelihood, the specifics of which are discussed in the next section. The set of output characters with the highest posterior probability is selected by the learner to fill in its blanks. In case of a tie, the set that is closer to the teacher's edge outputs is chosen. After the learner completes this inference process, it is ready to become a teacher. A new learner agent is created and the cycle repeats with the old learner as the new teacher. This process continues for a set number of generations.

Bayesian Inference Formulation

Bayesian inference has two primary components, the prior probability of a hypothesis, and the likelihood of the hypothesis given the data. There is also a third component, the marginal probability of the data. However, this component is constant and in the interest of simplification we will drop it in our calculations.

Our investigation of methods of calculating likelihood looks at three different measures. The first is a control likelihood that is always one, Equation (2). The second is a likelihood measure based on expressivity. Expressivity makes a plausible likelihood measure because the more distinct signals a hypothesized transducer is able to make the more likely that its signals can be decoded back into the correct meaning. Our measurement of expressivity looks at the list of output strings produced for the training input strings and simply divides the number of different strings by the total number of strings, Equation (3).

The third likelihood calculation is based on comprehension; how likely a transducer is able to decode, when reversed, its encodings of the training set. A hypothesis that results in a transducer that has this internal consistency is considered more likely. Essentially, an agent checks whether a hypothesized language allows the agent to talk to itself as in Mirolli and Parisi (2006). The likelihood for a given input-output mapping is calculated by counting the number of paths through the reversed transducer that read the output characters and write the correct input characters divided by the total number of paths that read the output characters. The likelihood is never zero because there is always at least one path that will write the correct characters. The final likelihood for the hypothesis is the average over all of the input-output mappings drawn from the input training strings. Equation (4) shows this calculation, with R being the set of training strings and $|R|$ the size of the training set. Each input training string is equally likely, so the average is not weighted.

$$P(H|D) = 1 \quad (2)$$

$$P(H|D) = \frac{DifferentOutputs}{TotalOutputs} \quad (3)$$

$$P(H|D) = \frac{\sum_{s \in R} \frac{SuccessfulDecodingPaths_s}{TotalDecodingPaths_s}}{|R|} \quad (4)$$

$$DL(Transducer) = N_{Edges} * (2 * \lceil log_2(N_{States}) \rceil) \quad (5)$$

Our prior calculation weights hypotheses by how much the resulting transducer can be compressed. The size of the transducer is measured as description length in bits by calculating the cost of storing each edge based on the number of states, Equation (5). The compressed size, DL_c, is compared to the size of the transducer before compression,

DL_u. Equation (6) shows the formula that calculates the prior such that the more a transducer is able to be compressed the higher the probability. $DL_u + 1$ is used in the calculation to ensure that the prior is never zero. A second control prior that is always one is also used, Equation (7).

$$P(H) = \frac{(DL_u + 1) - DL_c}{DL_u + 1} \quad (6)$$

$$P(H) = 1 \quad (7)$$

Results

We demonstrate the results of two experiments that investigate the generalization performance of the likelihood and prior measures. For each experiment, the input and output alphabets are both of size two. The length of every input string is 8 and consequently the length of every output string is 8. Each experiment has a training set of a specified size and the test set is all 256 strings of length 8, so the training set is a subset of the test set. Generalization performance is measured using the expressivity metric across the entire test set, rather than just the training set as it is used in the learning process.

Experiment One

The first experiment uses a training set of 16 input strings with one of the strings randomly chosen each generation to not have its corresponding output conveyed to the learner. This results in average of 3.3 blanks, with a standard deviation of 2.2, to be inferred by the learner out of total of 53.74 edges on average. The results shown here are the average expressivity across 50 trials each with a different randomly chosen training set. The experiment runs for 200 generations of teacher-learner interactions.

Figure 1 shows a plot of the expressivity over time, with standard deviation bars, using the description-length prior and each of the three likelihood measures: flat, expressivity-based and comprehension-based. We see that all three measures start with similar levels of expressivity but the comprehension measure quickly jumps ahead of the other two measures. It continues this rapid ascent before plateauing at slightly over 90% expressivity. The expressivity-based measure also ascends but much more slowly and settles in slightly above 26%. This isn't bad considering that the training set is only 6.25% of the test set, but it falls well short of success of the comprehension-based measure. The flat measure establishes a baseline that ends around 15%.

Figure 2 compares the expressivity when using the description-length prior versus the flat prior under the comprehension-based likelihood. The description-length prior results in clear improvement in expressivity. But, the flat prior turns in a respectable performance that ends at almost 70%.

Experiment One shows that agents that try to maximize comprehension are much better at generalizing than agents

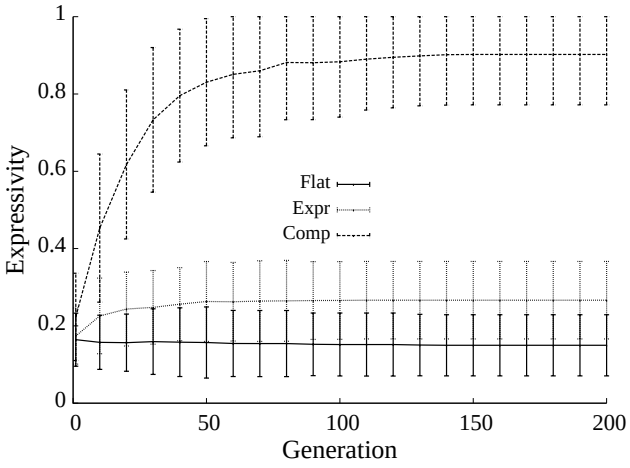

Figure 1: Likelihoods, 16 Training Strings

that try to maximize expressivity. The results from the analysis of the priors indicate that seeking to maximize compression in addition to maximizing comprehension results in even better generalization. The verdict on expressivity as a likelihood measure doesn't look good, but we want to make sure that the small training set isn't setting up expressivity to fail.

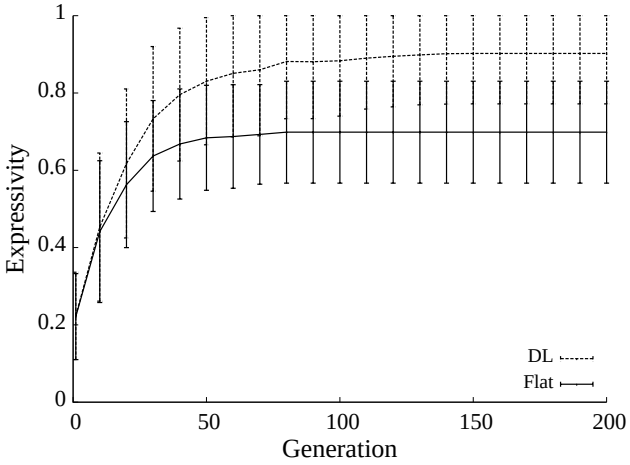

Figure 2: Priors, 16 Training Strings

Experiment Two

The second experiment uses a training set of 64 input strings, four times larger than the first experiment. Again, one of the strings is randomly chosen each generation to not have its corresponding output conveyed to the learner. The results shown here are the average expressivity across 50 trials each with a different randomly chosen training set.

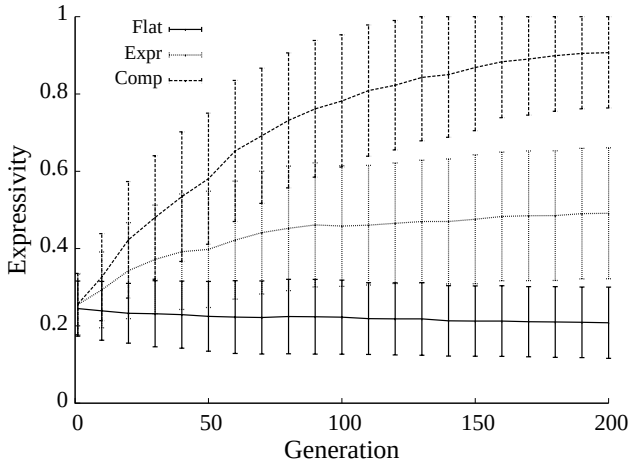

Figure 3: Likelihoods, 64 Training Strings

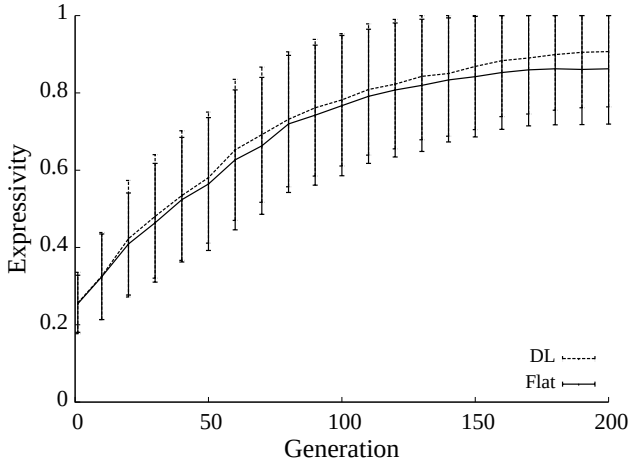

Figure 4: Priors, 64 Training Strings

Figure 3 shows the plot of the three likelihood measures. The comprehension-based measure is at the top ending again at slightly over 90%. The expressivity-based measure gets a boost from the larger training size and reaches slightly below 50%. The flat likelihood doesn't do much better than before settling in at around 20%. Once again, maximizing comprehension results in significantly better generalization than trying to maximize expressivity.

The analysis of the priors, using the comprehension-based likelihood, for Experiment Two is in Figure 4. Interestingly, we see that the that there isn't a significant benefit to maximizing compression with the larger training set. The training set is now large enough that seeking to maximize likelihood is sufficient to achieve high expressivity.

Conclusions

The capability to generalize is the hallmark of a compositional system. The Bayesian agents' ability to generalize their encodings to novel strings means that their communications are compositional. From the low expressivity at the start we can see that the compositionality emerges during training.

The success of the comprehension-based likelihood measure over the expressivity-based one demonstrates the value of including comprehension in the process. It is not sufficient to concentrate just on production and how many signals an agent can make. The pressure of being forced to actually decode those signals back into meanings is necessary to drive the emergence of a generalizable grammar.

The benefit of the description-length prior reaffirms the value of simplicity-based metrics like MDL. The added pressure to compress the grammar allowed the agents to express a large majority of the test set even with a very small training set. However, the prior's value decreases as the agents access more information. Large training sets mean that prior knowledge is no longer necessary to master the test set.

The iterated learning model again proves to be a powerful method of modeling the emergence of compositional grammars. The Bayesian version provides us with new ways of analyzing the process with the clear delineation of the role of the prior and the likelihood. The experiments here show that choosing a successful likelihood measure is not as simple as it might seem. A metric like expressivity seems like a good candidate but turns out to be rather poor. Likewise, the prior should be carefully chosen; a good prior can make the difference when knowledge is scarce. Finding two that work together, in this case the likelihood's pressure to be comprehensible and the prior's pressure to be simple, is the key to successful Bayesian inference and might be the key to our ability to generalize as well.

References

Batali, J. (1998). Computational simulations of the emergence of grammar. In Hurford, J., Studdert-Kennedy, M., and Knight, C., editors, *Approaches to the Evolution of Language: Social and Cognitive Bases*, pages 405–426. Cambridge University Press, Cambridge.

Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial Life*, 8(1):25–54.

Brighton, H. (2003). *Simplicity as a Driving Force in Linguistic Evolution*. PhD thesis, Theoretical and Applied Linguistics, The University of Edinburgh.

Brighton, H. (2005). Linguistic evolution and induction by minimum description length. In Werning, M. and Machery, E., editors, *The Compostionality of Concepts and Meanings: Applications to Linguistics, Psychology and Neuroscience*, pages 405–426. Ontos Verlag, Frankfurt.

Ferdinand, V. and Zuidema, W. (2009). Thomas' theorem meets bayes' rule: a model of the iterated learning of language. In *Proceedings of the 31th Annual Conference of the Cognitive Science Society*.

Griffiths, T. L. and Kalish, M. L. (2005). A bayesian view of language evolution by iterated learning. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*.

Griffiths, T. L. and Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, 31(3):441–480.

Kalish, M. L., Griffiths, T. L., and Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin and Review*, 14(2):281–294.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2):102–110.

Kirby, S. (2007). The evolution of meaning-space structure through iterated learning. In Lyon, C., Nehaniv, C., and Cangelosi, A., editors, *Emergence of Communication and Language*, pages 253–268. Springer Verlag.

Mirolli, M. and Parisi, D. (2006). Talking to oneself as a selective pressure for the emergence of language. In *Proceedings of the 6th International Conference on the Evolution of Language*, pages 214–221.

Rafferty, A., Griffiths, T., and Klein, D. (2009). Convergence bounds for language evolution by iterated learning. In *Proceedings of the 31th Annual Conference of the Cognitive Science Society*.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14:465–471.

Schrementi, G. and Gasser, M. (Forthcoming, 2010). Minimum description length and generalization in the evolution of language. In *Proceedings of the 8th International Conference on the Evolution of Language*.

Smith, K., Kirby, S., and Brighton, H. (2003). Iterated learning: a framework for the emergence of language. *Artificial Life*, 9(4):371–386.

Vitanyi, P. M. B. and Li, M. (2000). Minimum description length induction, bayesianism and kolmogorov complexity. *IEEE Transactions on Information Theory*, 46(2):446–464.

Wallace, C. S. and Boulton, D. M. (1968). An information measure for classification. *Computer Journal*, 11(2):185–194.