A Computational Model of Language Learning Driven by Training Inputs

Eun Seok Lee[1], Ji-Hoon Lee[2], Byoung-Tak Zhang[3]
Program of Cognitive Sciences[1], Program of Bioinformatics[2],
School of Computer Science and Engineering[3], Seoul National University
eslee@bi.snu.ac.kr

Abstract

Language learning involves linguistic environments around the learner. So the variation in training input to which the learner is exposed has been linked to their language learning. We explore how linguistic experiences can cause differences in learning linguistic structural features, as investigate in a probabilistic graphical model. We manipulate the amounts of training input, composed of natural linguistic data from animation videos for children, from holistic (one-word expression) to compositional (two- to six-word one) gradually. The recognition and generation of sentences are a "probabilistic" constraint satisfaction process which is based on massively parallel DNA chemistry. Random sentence generation tasks succeed when networks begin with limited sentential lengths and vocabulary sizes and gradually expand with larger ones, like children's cognitive development in learning. This model supports the suggestion that variations in early linguistic environments with developmental steps may be useful for facilitating language acquisition.

Introduction

One of the critical aspects of language learning is that it develops. Different from traditional computational approach which considers language learning as an innate rule learning and template matching, developmental accounts argue that infants' language learning depends on their environment, in particular the linguistic environment (Kaplan, Oudeyer, and Bergen 2008). We bring in developmental model methodology emphasizing computational learning in an incremental and open-ended way. Specifically, we explore the relationship between language capacity and its language environments (training inputs).

Elman (1993) showed that a gradual increase of attention span or, equivalently, a gradual increase of memory size allowed his neural networks to solve tasks that were unsolvable when starting with a `full-grown' network. Following Elman, we let the agents themselves go through developmental stages, and in addition to that, we manipulate the world (the amount of training input).

We consider a computer agent which takes a stream of various commercial video scripts for children step by step and progresses in language learning. Specifically, we investigate the use of the DNA hypernetwork model for learning to generate sentences based on a text collection of natural dialogues. Hypernetworks are originally proposed as an associative memory model inspired by and realized in molecular self-assembly (Zhang and Jang, 2006). A hypernetwork consists of a huge number of hyperedges, each of which links vertices of arbitrary size and thus is able to encode higher-order interactions or constraints among the variables. This view of hyperedges as constraints extends the application range of hypernetworks far beyond the associative memory (Chen et al., 2005) to associative processors.

Using hypernetwork structure with growing data from video corpus, it develops a concept for a given keyword by the associative memory organizing mechanism. Based on its plausibly crafted concept, we test its ability to generate sentences on the given keyword. We simulate the DNA hypernetworks to learn a language model and to generate new sentences based on a text corpus of approximately 30,000 sentences collected from animation videos for children with amount control. And we show the result of experiments in which concepts developed and generating sentences for a given keyword. By doing this, we check our expectation

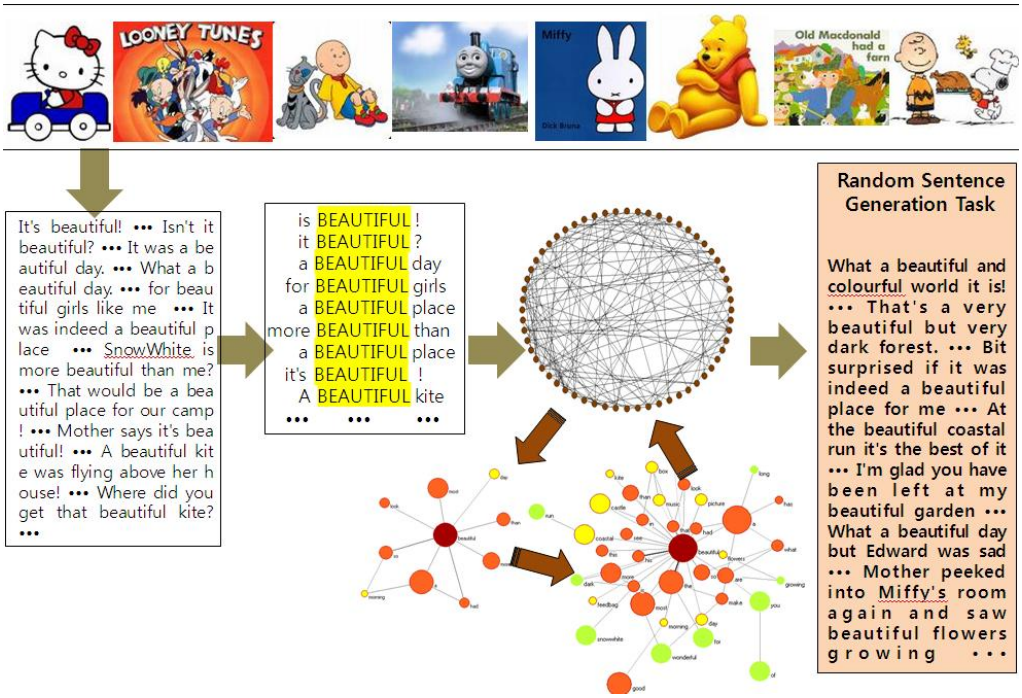

Figure 1. Process of generating a new sentence from a keyword (in this case, the keyword is "beautiful"). The given keyword is extended by assembling a new word to the left and right ends of the existing partial sentence.

that the hypernetwork be able to show some language learning capacity and its language environments be clearly and closely related with each other.

Learning Based on the Data

The experiments in this paper are based on commercial video scripts for children. For the training text corpus input, we used commercial and educational animation video scripts for children.* This corpus consists of S = 32,744 sentences excerpted from animation video scripts used in educational curricula from three to seven years old child. The script data are divided into 11 different learning stages classified by reading difficulty level. The corpus has 6,124 word types and 252,936 word tokens.

Consider a language learner which takes a stream of linguistic data. It interacts with linguistic data online, develops its initial concepts on some specific linguistic items, and has an internal representative semantic structure for given stimulus.

Figure 1 shows a high-level sketch of the complete model. The intuition behind this architecture is as follows.

The language learner takes a stream of various commercial video scripts for children step by step and progresses in language learning. A linguistic hypernetwork, the language learner, represents a probabilistic model of the data set using a population of hyperedges and their weights. See Figure 2 for an example of such a concept map, and how it deals the set of concepts associated which are insensible, but which gradually governs semantic coherence of its language.

The task is to learn a language model $P(S) = P(\mathbf{x}) = P(x_1, \ldots, x_n)$ from a collection of example sentences $D = \{\mathbf{x}^{(i)}\}$. Given a list of query words or a query sentence $\mathbf{x}^{(q)}$, the model is to generate a (potentially) new sentence $\mathbf{x}^{(a)}$.

To solve the language generation problem, we estimate the joint probability of words, $P(\mathbf{x}) = P(x_1, x_2, \ldots, x_n)$ as a language model. Given a query sentence with the i-

* The titles of the video materials are as follows: *Miffy, Looney the Tune, Caillou, Dora Dora, McDonald, Timothy, Kitty, Snoopy.* The learning order fixation of materials is based on the recommended consumer ages of the each video product.

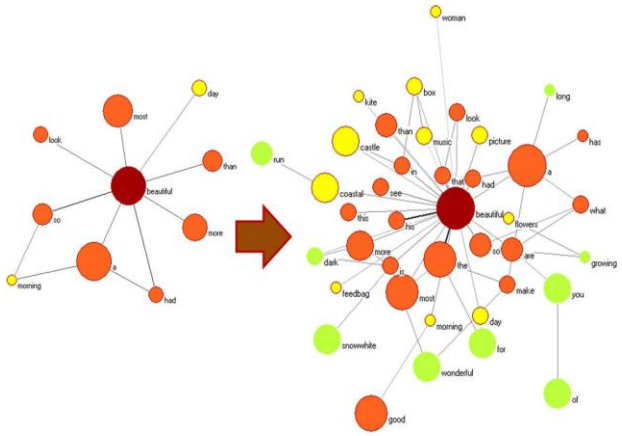

Figure 2. A concept for a word 'beautiful' extracted from a linguistic environment and the typical development of the concept for the given keyword 'beautiful' through hypernetwork. It has connections to other words and capacities to expand these connections.

th position blank, $\mathbf{x}_{-\mathrm{i}}^{(q)}=(x_1^{(q)}, x_2^{(q)}, \ldots, x_{\mathrm{i}-1}^{(q)}, x_{\mathrm{i}+1}^{(q)}, \ldots, x_n^{(q)})$ as context or history h, the model is used to choose the word x^* as

$$x_{i*} = \arg\max_{x_i} P\left(x_i \middle| \mathbf{x}_{-\mathrm{i}}^{(q)}\right) = \arg\max_{x_i} P(x_i|h)$$

where x_{i*} is the word that maximizes the conditional probability.

Conventional statistical language models estimate the probability of a sentence S by using the chain rule to decompose it into a product of conditional probabilities:

$$P(S) = P(\mathbf{x}) = P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid x_1, \ldots, x_{i-1})$$

$$= \prod_{i=1}^{n} P(x_i \mid h)$$

where $h = \{x_1, \ldots, x_{i-1}\}$ is the history when predicting word x_i. Maximum entropy (ME) models have been successfully applied in language modeling to approximate conditional probabilities of the form $P(x \mid h)$, as described in (Rosenfeld, 1996).

Maximum entropy models are exponential distributions which satisfy given linear constraints. In conditional language modeling, given constraints or "feature" functions $f_i(h, x)$ and respective desired values, the ME solution is expressed as

$$P(x|h) = \frac{1}{Z(h)} \exp\left(\sum_i w_i f_i(h, x)\right)$$

The parameters w_i can be derived using the generalized iterative scaling algorithm (Rosenfeld, 1996 and references therein). This word-based model can be generalized to a sentence-based ME model:

$$P(S) = \frac{1}{Z} \exp\left(\sum_i w_i f_i(S)\right)$$

by mapping $x \to S$ and $h \to \epsilon$ (the null history). Many different types of features or constraints have been proposed and implemented based on the conventional *N*-gram models and those that have been used in conditional ME models.

This idea can be implemented straightforwardly by the DNA hypernetwork architecture (Zhang, 2008). To see this, we reformulate the sentence-based ME model as

$$P(S) = P(x_1, \ldots, x_n) = \frac{1}{Z} \exp\left(\sum_i w_i f_i(x_1, \ldots, x_n)\right)$$

$$= \frac{1}{Z} \exp\left(\sum_{k=1}^{K} \frac{1}{C(k)} \sum_{i_1, i_2, \ldots, i_k} w_{i_1 i_2 \ldots i_k}^{(k)} x_{i_1} x_{i_2} \ldots x_{i_k}\right)$$

where $w_{i_1 i_2 \ldots i_k}^{(k)} x_{i_1} x_{i_2} \ldots x_{i_k}$ is a hyperedge with weight, and $K \leq n$. The feature functions, i.e. hyperedges, are defined as combinations of arbitrary k words chosen from $x_1, \ldots, x_n$. Note that the newly defined function structure is not restricted to the *n*-gram models, i.e. the "hyperfeatures" can be an arbitrary number of non-contiguous words. We remind that this "linguistic" hypernetwork is a weighted hypergraph, $H = (V, E, W)$, where V is the set of vertices representing the words and E is the set of hyperedges representing the phrases (hyperedges are edges that can connect an arbitrary number of vertices), and W is the weight of the hyperedge.

Simulation Experiments

To evaluate the potential of language generation using the DNA hypernetworks we experimented with the

following procedure. In this scenario, the query is given as a single word which is then extended bidirectionally, left and right simultaneously. Alternatively, we can extend the words only to the right, which is more like humans compose sentences in countries where people read and write left to right. The sentence generation procedure is summarized as follows.

Step 1. Given a keyword L_q=(x_q), retrieve the hyperedges into M=$\{L_1,L_2,\ldots,L_m\}$.

Step 2. Select a hyperedge L_h=(x_{q-1}, x_q, x_{q+1}) from M using roulette wheel selection.

Step 3. Set L_q = (x_{q-1}, x_q) and do Steps 1 and 2. Set the resulting hyperedge L_{left}=(x_{q-2}, x_{q-1}, x_q)

Step 4. Set L_q = (x_q, x_{q+1}) and do Steps 1 and 2. Set the resulting hyperedge L_{right}=(x_q, x_{q+1}, x_{q+2})

Step 5. Generate a (partial) sentence L=(x_{q-2}, x_{q-1}, x_q, x_{q+1}, x_{q+2}) by combining L_{left} and L_{right}.

Step 6. Repeat Steps 3-5 by extending L_q until the termination condition for sentence generation is met (see text).

After learning, we examine the agent's linguistic structure for a given keyword. Figure 2 shows a concept for a keyword 'beautiful' expands as learning proceeds. The word associations with the keyword show that the agent develops its concept into well-structured one according to training inputs. Interestingly, we can find that associated words are both syntactically and semantically well suited for the given word. For example, we can find suitable syntactic markers such as 'a,' 'the,' 'more' grows apparently. And in the aspect of semantics, there is also some coherence within it.

In addition to the concept development, we analyzed 100 sentences generated by the agent on given keyword at each learning step. The agent generates random sentences on a given keyword (i.e. beautiful) in both ways: developmental learning and improvised learning. For example, apparently the sentence "beautiful hello lucky" is not coherent. But the sentence "what a beautiful world" is coherent, in the aspect of syntax and semantics. In the case of developmental learning, the input to the agent is incremental and developmental, so that the agent is exposed to the input which takes in previous learning stages. In contrast, improvised learning has no contribution from the past. Table 1 shows the excerpts of well-structured sentences generated with developmental learning ways. In earlier learning stages, it produces relatively shorter sentences with well structure. As learning develops, it generates more complex structured ones. Figure 3 shows the typical proportion of well-structured sentences after learning in both ways at the simulation experiments. In the developmental learning ways, ass more training data comes into the network, it generates more coherent sentences. Contrary to that, improvised learning doesn't show such features.

Earlier learning stage	Later learning stage
A beautiful world	What a beautiful and colourful world it is!
In a very dark forest	That's a very beautiful but very dark forest.
A beautiful place for me	Bit surprised if it was indeed a beautiful place for me
The beautiful costal run	At the beautiful coastal run it's the best of it
My beautiful garden	I'm glad you have been left at my beautiful garden
A beautiful day	What a beautiful day but Edward was sad
Beautiful flowers	Mother peeked into Miffy's room again and saw beautiful flowers growing

Table 1. Excerpted sentences generated from earlier learning stages to later ones with developmental learning

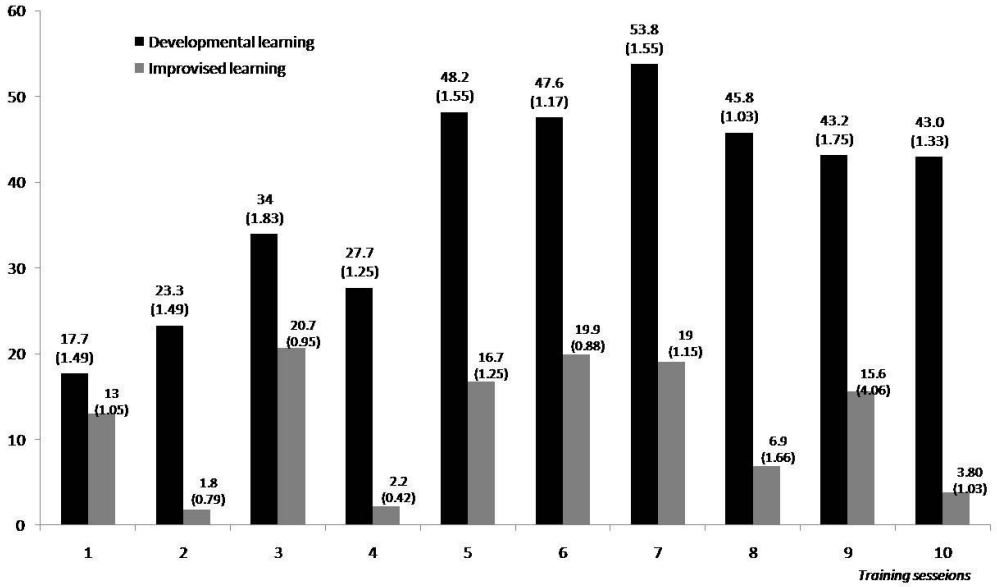

Figure 3. The typical proportion (%) and *SD* of well-structured sentences from random sentences generation task with developmental learning and improvised learning.

Conclusion

In this paper, we explore how linguistic experiences can cause differences in learning linguistic structural features, as investigate in a probabilistic graphical model.

We simulated a computer agent which takes a stream of various commercial video scripts for children step by step and progresses in language learning. Specifically, we used the DNA hypernetwork model for learning to generate sentences based on a text collection of natural dialogues. Hypernetwork is discussed as a framework for concept-driven and developmental language learning.

The simulation results show that language learning from the statistical distribution of large sets of linguistic data may be a nontrivial source of linguistic communication, and hypernetwork may be useful for growing representational structures and facilitate language acquisition.

Simulation results focused on the computational properties of the model, with the goal of showing that incrementally injected training data leads to reliable language learning. The experiments involve testing performance of models with some aspect of the data driven language learning. Although our investigation of this architecture is just beginning, we have shown that the model can explain some fundamental behavioral data. As illustrated in this research, computational approaches may shed new light on the particular role played by language learning mechanisms in complex linguistic developmental processes. We have presented an experimental setup to explore a scenario on the basis of the hypothesis that language learner may discover communication through a general process of probabilistic and developmental associative memory. However, it should be clear that we do not suggest that developmental learning progress is the only motivational principle driving children during their development. Development certainly results from the interplay between a complex set of drives, particular learning biases, as well as embodiment and environmental constraints. Our hope is that this form of experiment can help to develop our intuitions and to better understand the different components that contribute to shaping the dynamics of child's linguistic development.

References

Elman, J.L. (1993). Learning and development in neural networks: The importance of starting small. Cognition, 48, 71-99.

Elman, J.L. (2005). Connectionist models of cognitive development: Where next? Trends in Cognitive Science, 9, 111-117.

Steels, L., (2005) 'The Emergence and Evolution of Linguistic Structure: From Lexical to Grammatical Communication Systems', Connection Science, 17(2), 213–230.

Borovsky, A., & Elman, J.L. (2006). Language input and semantic categories: A relation between cognition and early word learning. Journal of Child Language, 33, 759-790.

Vogt, P. (2005). On the acquisition and evolution of compositional languages: Sparse input and the productive creativity of children. Adaptive Behavior (Special issue on Evolution and Acquisition of Language), 13(4).

Kaplan, F., Oudeyer, P-Y., Bergen B., (2008) 'Computational models in the debate over language learnability', Infant and Child Development, 17(1), 55–80.

Chomsky, N. (1975). Reflections on language. New York: Pantheon Books.

Smith, K., Brighton, H., & Kirby, S. (2003). Complex systems in language evolution: the cultural emergence of compositional structure. Artificial Life, 9(4), 371-386.

Keibel, H., Elman, J. L., Lieven, E. & Tomasello, M. (2005). From words to categories : distributional regularities in German child-directed speech.

Hart, B. & Risley, T. (1995). Meaningful differences in the everyday experience of young American children. Baltimore, MD: Paul H Brookes Publishing CO, 1995.

Rosenfeld, R., (2000) "Two decades of statistical language modeling: where do we go from here?", *Proceedings of the IEEE*, **88(**8).

Rosenfeld, R., (1996) "A maximum entropy approach to adaptive statistical language modeling," *Computer Speech and Language*, **10**:187-228.

Zhang. B.-T. and Jang, H.-Y. (2006). Molecular learning of wDNF formulae, Lecture Notes in Computer Science, *DNA 11*, 3892:427-437.

Zhang, B.-T. (2008). Hypernetworks: a molecular evolutionary architecture for cognitive learning and memory. *IEEE Computational Intelligence Magazine*, 3(3), 49-63.

Zhang, B.-T. and Park, C.-H. (2008) "Self-assembling hypernetworks for cognitive learning of linguistic memory," *Proc. Int. Conf. Comp., Elect., and Syst. Sci., and Eng.* vol. 27, pp.134-138, 2008.