Using Semantic Information for Language Learning

Dana Angluin and Leonor Becerra-Bonache
Department of Computer Science, Yale University, USA
{dana.angluin, leonor.becerra-bonache}@yale.edu

Formal language learning models can provide useful insights into the developmental stages by which children learn their native language. Several questions in natural language learning may be addressed by studying these formal models, e.g., the impact of semantic information on learning the syntax of a language and the kind of information available to the learner. Moreover, these formal models are also motivated by the practical applications of language learning by machines [2]. With the current work we hope to contribute a deeper understanding of the role of semantics in language acquisition.

Computational approaches to children's language acquisition tend to omit semantic information and reduce the learning problem to syntax learning. However, as Feldman states, "[...] no one believes that children learn the grammar of their native language independent of meaning (semantics) and use (pragmatics)" [1]. Semantics and context seem to play an especially important role in the 2-word stage of child linguistic development. Despite the substantial difference between their grammars, communication is possible between the child and the adult because they share a situation, and the meanings of elements of their utterances often indicate their implied syntactic relations [3].

Based on these ideas, we present a simple computational model that takes into account semantics for language learning. We focus initially on a simple formal framework, which we intend to develop into one with more cognitive plausibility. We consider a specific domain of geometric shapes and their properties and relations. The model accommodates two different tasks: comprehension and production; here, we focus only on comprehension.

We propose a formal model of meaning and denotation using finite-state transducers. Meaning and denotation functions are used by the teacher to provide examples for the learner. For instance, a meaning transducer for a class of sentences in English is given in Figure 1. The meaning assigned by this meaning transducer to the utterance *the triangle to the left of the red circle* is $< tr(x_1), le(x_1, x_2), re(x_2), ci(x_2) >$.

Situations, formalized as sets of ground atoms, represent the objects, properties and binary relations that are noticed in some environment of the teacher and learner. For example, noticing a big blue triangle above a big green square gives the following situation: $\{bl(t_1), bi(t_1), tr(t_1), ab(t_1, t_2), gr(t_2), bi(t_2), sq(t_2)\}$. To determine the denotation of an utterance u in a situation S, the teacher matches the meaning of u to a subset of S.

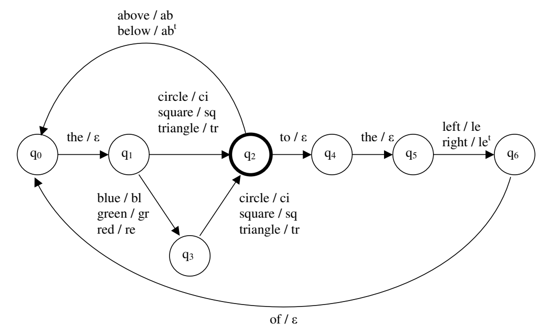

Fig. 1. A meaning transducer

In order to learn the meaning function, we assume that the learner receives a sequence of pairs (S_i, u_i) from the teacher, where S_i is a situation, and u_i is an utterance with a denotation in the situation S_i. Note that the meaning of u_i is not specifically isolated for the learner in S_i. The basic strategy used by the learner is cross-situational conjunctive learning [4], [5]. For each encountered word w, the learner considers all utterances u_i containing w and their corresponding situations S_i, and forms the intersection of the sets of predicates occurring in these S_i. Binary predicates are processed using the results for unary predicates.

We prove that our learning algorithm finitely converges to a correct meaning function under a specific set of assumptions about the transducer and examples. We test our model with sets of utterances in a number of natural languages, including Arabic, English, Greek, Hebrew, Hindi, Mandarin, Russian, Spanish and Turkish, and use the results to illustrate the strengths and limitations of our approach.

REFERENCES

[1] J.A. Feldman. Real language learning. In V. Honavar and G. Slutzki, editors, *4th International Colloquium on Grammatical Inference*, pages 114–125. Springer, 1998.

[2] R. Parekh and V. Honavar. Grammar inference, automata induction and language acquisition. In H. Moisl R. Dale and A. Somers, editors, *Handbook of Natural Language Processing*, pages 727–774. Marcel Dekker, New York, 2000.

[3] A. M. Schaerlaekens. *The two-word sentence in child language development*. Mouton, La Haya, 1973.

[4] J.M. Siskind. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61:39–61, 1996.

[5] K. Smith, A.D.M. Smith, R.A. Blythe, and P. Vogt. Cross-situational learning: a mathematical approach. In P. Vogt, Y. Sugita, E. Tuci, and C. Nehaniv, editors, *Emergence and Evolution of Linguistic Communication III*, pages 31–44. Springer, 2006.