Learning more expressive local substitutable grammars.

Internship subject 2017-2018

Subject:

Learning grammars automatically was originally motivated by the problem of natural languages acquisition and is also of interest for other areas dealing with sequences, such as genomics, document processing, software engineering, robotics or even the detection of intrusions in a system...

The main recent breakthroughs in the field have taken roots in the framework of distributional learning, focusing on the contexts of the words. A contribution to these advances in our team has been the proposition of a class of substitutable grammars, defined with respect to local contexts, and a simple learning algorithm that have been shown to have nice theoretical properties and promising on real datasets experiments [1,2,3,4].

Building on this work, the subject of this internship is to propose a learning algorithm to learn more expressive local substitutable grammars (enabling for instance to restrict substitutability according to contextual information [5] and/or to model crossing correlations in the sequences [6]). Validation of the new learning approach will be theoretical (learnability or complexity results) or practical (testing its performances on artificial data or real enzymatic proteins thanks to our close collaborations with biologists).

Keywords: Machine learning, Formal Grammars, Algorithmics, Learnability

Required and appreciated skills: Basics in formal langage theory, algorithmics and machine learning (SML module is recommended).

Team: Dyliss, IRISA / Inria Rennes-Bretagne Atlantique

Advisor: François Coste, francois.coste@inria.fr, tel (33|0) 299 847 491

Bibliography:

1. Local Substitutability for Sequence Generalization, F. Coste, G. Garet and J. Nicolas, ICGI 2012, Washington.
   http://jmlr.csail.mit.edu/proceedings/papers/v21/coste12a/coste12a.pdf

2. A bottom-up efficient algorithm learning substitutable languages from positive examples, F. Coste, G. Garet and J. Nicolas, ICGI 2014, Kyoto.
   http://jmlr.csail.mit.edu/proceedings/papers/v34/coste14a.pdf

3. Classification and Characterization of Enzymatic Families using Formal Methods, G. Garet, Ph.D. thesis, Université de Rennes 1, Décembre 2014
   http://www.irisa.fr/dyliss/public/ggaret/these_Gaelle_Garet.pdf

4. Efficiently Learning (Local) Substitutable Context-Free Languages from Sequences by Grammar Reduction, F. Coste, G. Garet and J. Nicolas, in preparation.

5. A Refined Parsing Graph Approach to Learn Smaller Contextually Substitutable Grammars With Less Data, F. Coste, M. Demirdelen, ICGI 2016
   https://hal.inria.fr/hal-01406337

6. Distributional learning of parallel multiple context-free grammars, A. Clark, R. Yoshinaka, Machine Learning, 2014