A refined parsing graph approach to learn smaller contextually substitutable grammars with less data

François Coste, Mikail Demirdelen

Journal of Machine Learning Research () Submitted ; Published

A refined parsing graph approach to learn smaller contextually substitutable grammars with less data

François Coste FRANCOIS.COSTE@INRIA.FR
Mikail Demirdelen MIKAIL.DEMIRDELEN@IRISA.FR
Dyliss project, INRIA, Campus de Beaulieu, Rennes, France
UMR 6074 IRISA, CNRS, Campus de Beaulieu, Rennes, France

Editor:

Abstract

We present a refined parsing graph approach to learn smaller contextually substitutable grammars from smaller training samples in the framework initiated by ReGLiS algorithm.

Keywords: Inference, Substitutable languages, Parsing graph.

1. Introduction

Since the seminal work on learning substitutable languages by Clark and Eyraud (2007), new learnable classes of languages based on substitutability principle have been defined. Yoshinaka (2008) extended substitutable languages to define k,l-substitutable languages by restricting substitutability to occur only contextually in subcontexts of lengths k and l. Motivated by the application to proteins, Coste et al. (2012a) introduced the k,l-local substitutable languages requiring only local contexts of length k and l, rather than global contexts, to define substitutability classes. Mixing local and contextual substitutability enabled them to define the classes of k,l-local-contextually substitutable languages.

Learning such languages was initially done with variants (Yoshinaka, 2008; Coste et al., 2012a) of the SGL algorithm from Clark and Eyraud (2007). To tackle practical applications, Coste et al. (2012b, 2014) presented ReGLiS, an efficient learning algorithm based on dynamic programming techniques over a parsing graph structure. Initially designed for local substitutable and substitutable languages, ReGLiS builds a parsing graph using the substitutability congruence classes detected in the training sample. Following the SGL extension proposed by Yoshinaka (2008), the straightforward extension to contextually substitutable languages (the k,l-substitutable and k,l-local-contextually substitutable languages) is to use the congruence classes $[uyv]$ where $\langle u,v\rangle$ is a required k,l-context $\langle u,v\rangle$ restricting contextually the substitutability of y with others y'. As shown by the example in Figure 1, this approach is not optimal if contexts overlap each other: in that case, only one of the congruence class can be used if we keep the original algorithm while intuitively we'd like to use both. We denote by $N_y^{\langle u,v\rangle}$ the set of strings y' substitutable with y in context $\langle u,v\rangle$: we have $N_y^{\langle u,v\rangle} = \{y' \in \Sigma^+ : uy'v \in [uyv]\}$. Considering only arcs for each $N_y^{\langle u,v\rangle}$ could lead to substitutions modifying the required context. We propose in algorithm 1 a refined minimal parsing graph procedure based on a guard on the right of substitutability classes enabling to preserve required right and left contexts (see Demirdelen (2016) for details).

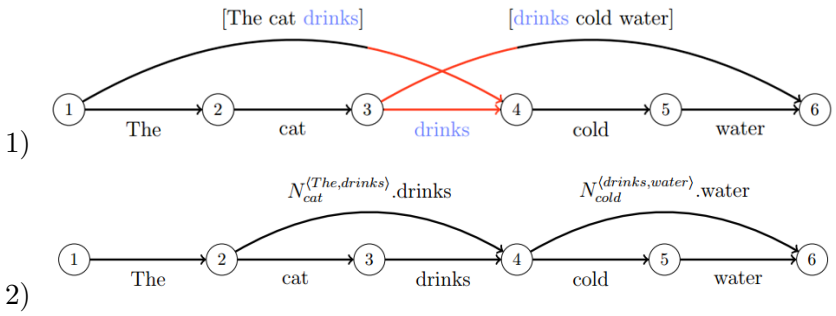

Figure 1: Parsing graph for the reduction of 'The cat drinks cold water' assuming $1,1$-contextual congruence classes [The cat drinks] and [drinks cold water]: 1) Based on congruence classes 2) Our solution using substitutability classes with guards.

Input: Parsed string α, language parameters k and l, set of substitutability classes $\mathcal{N}_\mathcal{C}$
Output: Set R of all non-redundant right-hand sides generating α

`/* Building parsing graph */`
$V \leftarrow \{i : i \in [1, |\alpha| + 1]\}$`/* vertex for each parsing position */`
$E \leftarrow \{(i, i + 1, N_{\alpha[i,i+1[}) : i \in [1, |\alpha|]\}$`/* symbol parsing edges */`
$d \leftarrow \max(k, l)$ `/* inter substitutability minimal distance */`
$E \leftarrow E \cup \{(i + k, (j - l) + d + 1, N^{u,v}.N_{\alpha[j-l,(j-l)+d+1[}) : i, j \in V, u.N^{u,v}.v = \alpha[i, j]\}$`/* parsing edges for each non-terminal and their minimal substitutability distance */`
`/* Searching for irreducible paths */`
$IP[1] \leftarrow \{'1'\}$ `/* Starting irreducible path */`
for $j \leftarrow 2$ ***to*** $|\alpha| + 1$ **do**
  $P \leftarrow \bigcup_{(i,j,l)\in E}\{p.l : p \in IP[i]\}$
  `/* Irreducible paths are the ones that have no other paths as subsequence */`
  $IP[j] \leftarrow \{p \in P : \nexists p' \in P, is\_subsequence(p', p)\}$
`/* Building right-hand sides from indices */`
$R \leftarrow \emptyset$
foreach $p \in IP[|\alpha| + 1]$ **do**
  $rhs \leftarrow \lambda$
  for $i \leftarrow 1$ ***to*** $|p| - 1$ **do**
    $rhs \leftarrow (rhs.l : (i, i + 1, l) \in E)$
  $R \leftarrow R \cup rhs$
return R

Algorithm 1: Refined Reduced_rhs for k, l-local contextually substitutability languages

References

A. Clark and R. Eyraud. Polynomial identification in the limit of substitutable context-free languages. *Journal of Machine Learning Research*, 8:1725–1745, August 2007.

F. Coste, G. Garet, and J. Nicolas. Local substitutability for sequence generalization. In *ICGI*, pages 97–111, 2012a.

F. Coste, G. Garet, and J. Nicolas. Learning context free grammars on proteins by local substitutability. unpublished first submission of ReGLiS, 2012b.

F. Coste, G. Garet, and J. Nicolas. A bottom-up efficient algorithm learning substitutable languages from positive examples. In *ICGI*, pages 49–63, 2014.

M. Demirdelen. Fast parser for biological sequences and a new algorithm for the inference of substitutable languages. Master's thesis, Master Recherche en Informatique, Université de Rennes 1, 2016.

R. Yoshinaka. Identification in the limit of (k,l)-substitutable context-free languages. In *ICGI*, pages 266–279, 2008.