Subregular Induction of Underlying Representations and a Phonological Grammar

Wenyue Hua, Adam Jardine, and Huteng Dai
{wenyue.hua,adam.jardine,huteng.dai}@rutgers.edu
Department of Linguistics, Rutgers University

Goals and Background

- Project goal: the **simultaneous inference** of URs and a grammar from SRs in a morphological paradigm. (Albright, 2002; Tesar, 2014)
- The **Input Strictly Local** (ISL) functions provide a structure that can solve this problem. (Chandlee and Heinz, 2018)

Primary result

- The learner induces UR and phonological grammar from a range of ISL_2 function (ISL function for $k=2$), including progressive and regressive assimilation, deletion, epenthesis, and opacity.

Learning Problem

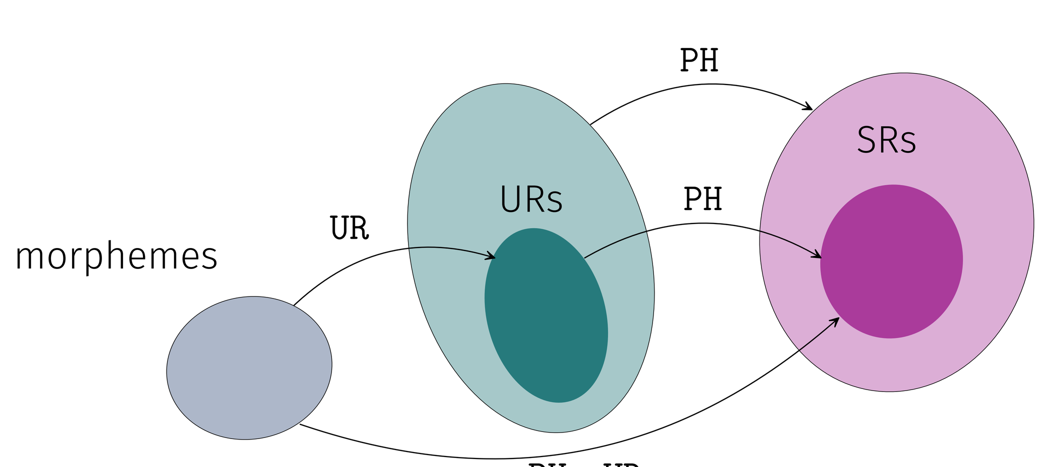
- M: finite set of morphemes {CAT, DOG, ..., PL}
  Σ: finite set of segments {a, b, β, ..., z}
- UR function: maps one morpheme to one UR;

$$\mathtt{UR}: M^* \rightarrow \Sigma^*$$

- PH function: maps URs to SRs;

$$\mathtt{PH}: \Sigma^* \rightarrow \Sigma^*$$

$$\begin{array}{llll} \mathtt{UR}(\text{CAT}) & = \text{kæt} & \mathtt{PH}(\text{kæt}) & = \text{kæt} \\ \mathtt{UR}(\text{PL}) & = \text{z} & \mathtt{PH}(\text{dɔgz}) & = \text{dɔgz} \\ \mathtt{UR}(\text{CAT-PL}) & = \text{kætz} & \mathtt{PH}(\text{kætz}) & = \text{kæts} \\ \ldots & \ldots & \mathtt{PH}(\text{bʊkz}) & = \text{bʊks} \\ & & \ldots & \ldots \end{array}$$



- Given a finite sample of $\mathtt{PH} \circ \mathtt{UR}$, how do we identify PH and UR?

(CAT-PL, kæts), (DOG-PL, dɔgz), ..., (BOOK-PL, bʊks)

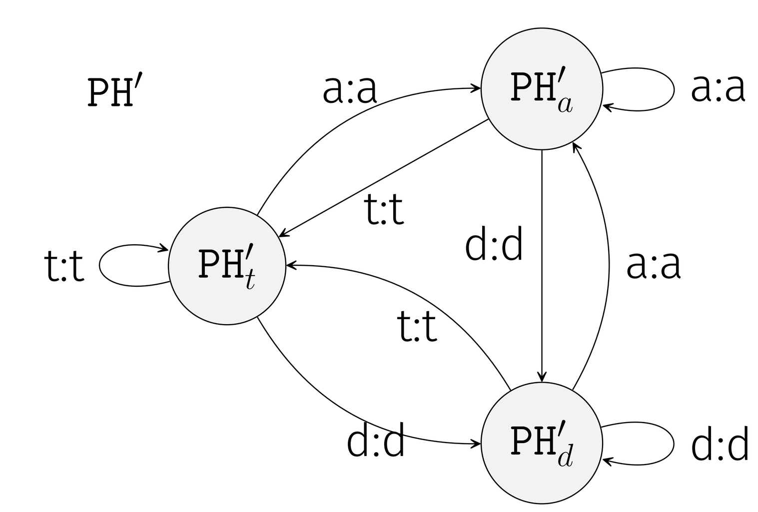
Initialization

- Running example $D \subset \mathtt{PH} \circ \mathtt{UR}(M^*)$

Sample of $\mathtt{PH} \circ \mathtt{UR}$ (PROG. ASSIMILATION)					
w	$\mathtt{PH}(\mathtt{UR}(w))$	w	$\mathtt{PH}(\mathtt{UR}(w))$	w	$\mathtt{PH}(\mathtt{UR}(w))$
r_1s_1	tatta	r_2s_1	tadda	r_3s_1	ata
r_1s_2	tatda	r_2s_2	tadda	r_3s_2	ada
r_1s_3	tata	r_2s_3	tada	r_3s_3	aa

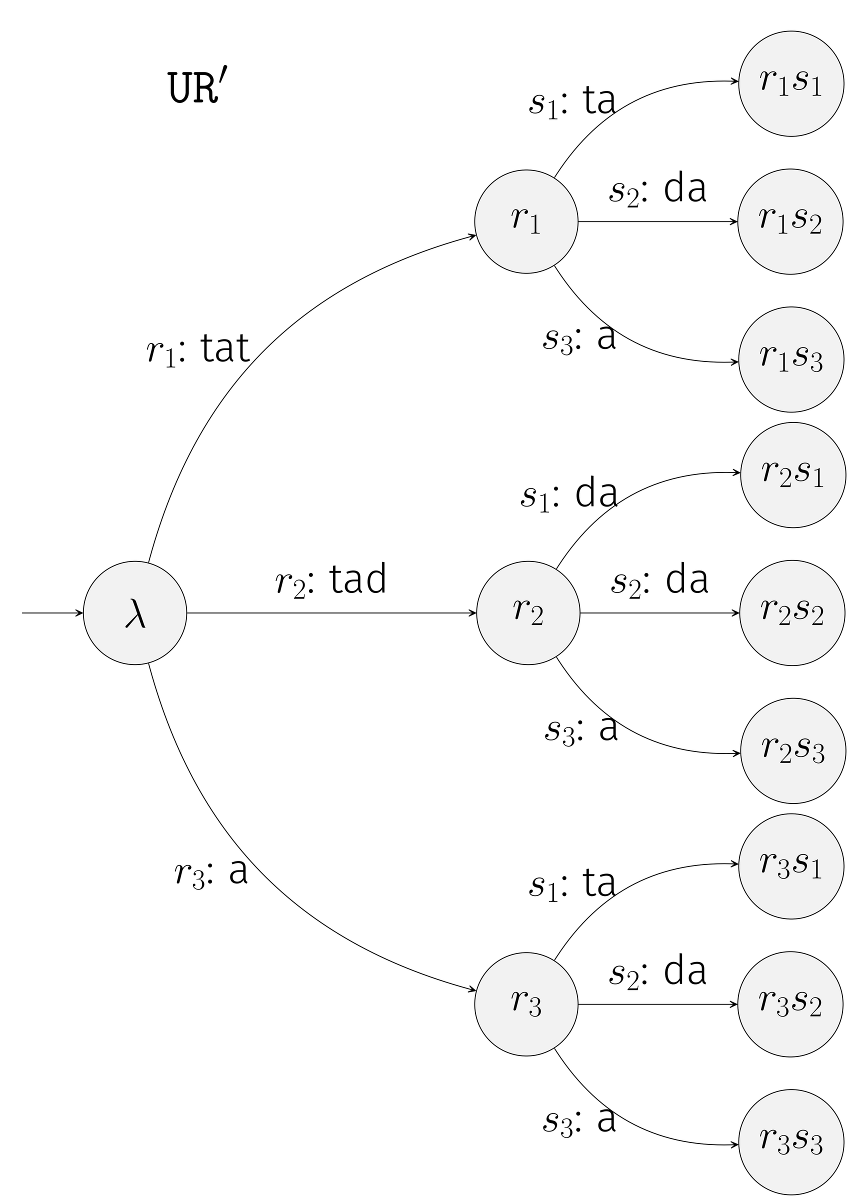
- Initialize PH′ to the identity function

$\mathtt{PH}'(\text{tatta}) = \text{tatta}$, $\mathtt{PH}'(\text{tadta}) = \text{tadta}$, etc.



- Initialize UR′ to a **prefix tree transducer** representing D: segmentation based on **longest common prefix (lcp)**.

$$\mathtt{lcp}(\{\underline{tat}ta, \underline{tat}da, \underline{tat}a\}) = tat$$
$$\mathtt{lcp}(\{\underline{tad}da, \underline{tad}a\}) = tad$$



Inconsistency detection

If a morpheme is mapped to multiple SRs, the learner detects this inconsistency.

r_1: tat r_2: tad r_3: a
s_1: ta, da s_2: da s_3: a

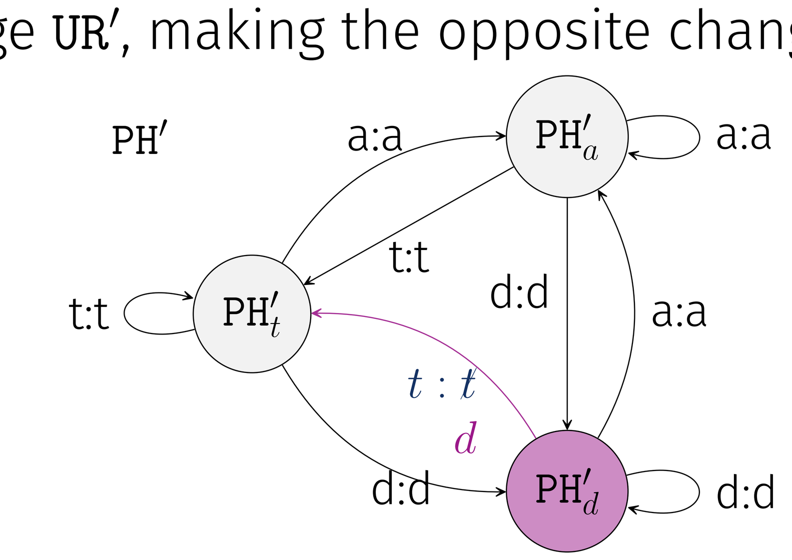
Environment collection

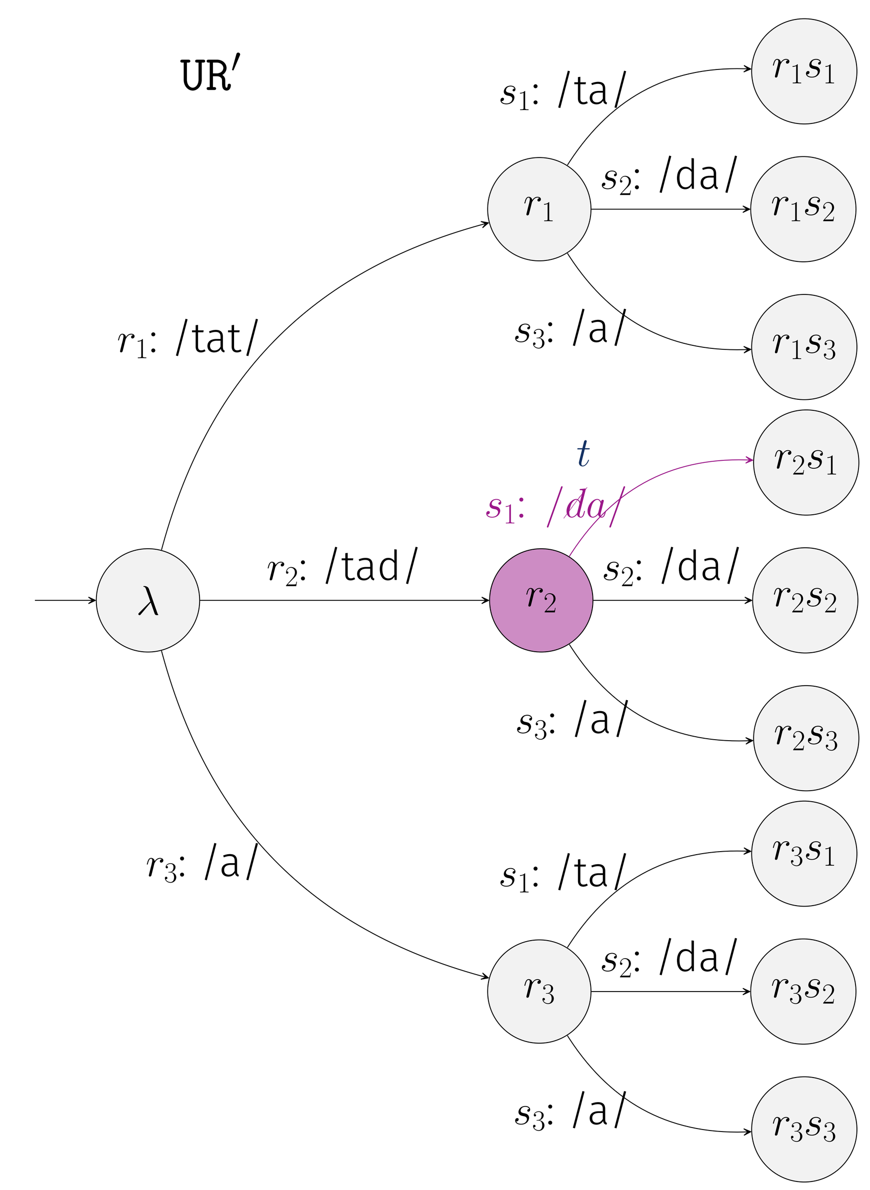
- ISL provides environment information
- t is the most **informative** form → UR.

ws_1	env. s_1
r_1s_1	tat ta
r_3s_1	a ta
r_2s_1	tad da

Modification

Change UR′, making the opposite change in PH′





Take-home message

From an **abstract** and **principled** perspective, learning is possible given the basic principles:

- a restrictive, structured hypothesis space
- complementarily distributed allomorphs
- a surface-driven set of URs
- One morpheme → one UR

Future work

- Long-distance processes can be captured by different classes of subsequential functions with a similar structure;
- One example: output strictly-local class also has a restricted state structure; (Chandlee et al., 2015)
- Abstract URs may be learnable when input alphabet is larger than output alphabet (and thus allows larger categories).

Selected References

Albright, Adam C (2002). *The identification of bases in morphological paradigms*. PhD thesis, University of California, Los Angeles.

Chandlee, Jane, Eyraud, Rémi, and Heinz, Jeffrey (2015). Output strictly local functions. In Kornai, Andras and Kuhlmann, Marco, editors, *Proceedings of the 14th Meeting on the Mathematics of Language (MoL 14)*, pages 52–63, Chicago, IL.

Chandlee, Jane and Heinz, Jeffrey (2018). Strict locality and phonological maps. *Linguistic Inquiry*, 49(1):23–60.

Tesar, Bruce (2014). *Output-driven phonology: Theory and learning*. Cambridge University Press.

Acknowledgements: We thank attendees of the Rutgers MathLing group, NECPhon 2019, and in particular Jeff Heinz, Charles Reiss, Bruce Tesar, Adam McCollum, and Colin Wilson for their insightful comments.

Take a picture to download the poster