Don't Stop Me Now!
Using Global Dynamic Oracles to Correct Training Biases of Transition-Based Dependency Parsers

Lauriane Aufrant[1,2], **Guillaume Wisniewski**[1] & **François Yvon**[1]

[1]LIMSI-CNRS, Univ. Paris-Sud, Université Paris-Saclay, France [2]DGA, France

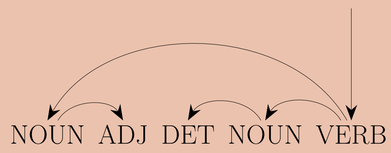
first.last@limsi.fr

Contributions

- An extended formalism, **dynamic** oracles for **global training** & beam search
  ↪ A **sound** error criterion, that allows eg. to
  - Train on **partial** data with full beam search consistency
  - Design many new **sampling/update strategies**
  - **Transpose** those already designed for local training
  - Replace both paradigms in the **same framework**
- A new global training strategy, that corrects several **sampling biases**

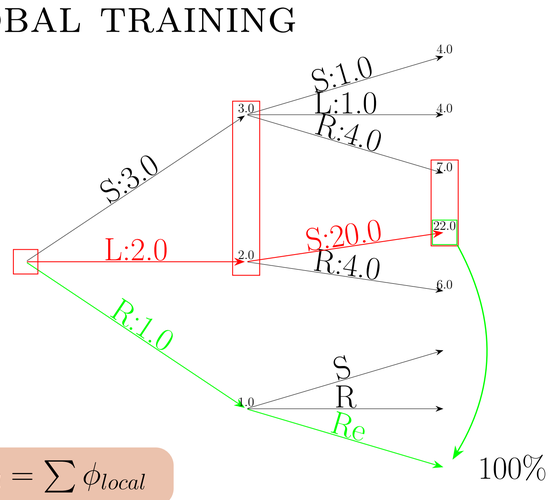
Focus on the **generation of candidate** configurations for model updates

Transition-based dependency parsing



Transitions operating on a stack and a buffer:
Shift, Left, Right, Reduce

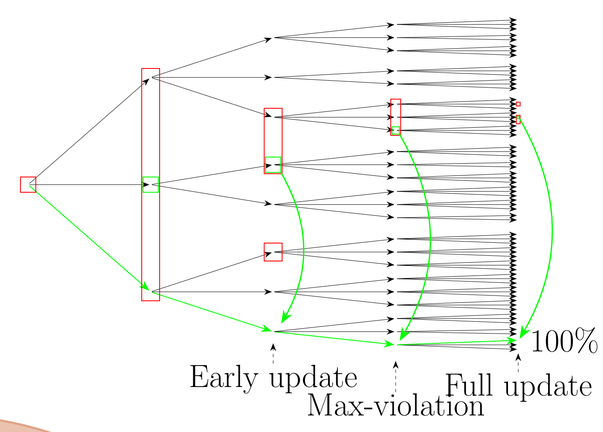
Global training



$\Phi_{global} = \sum \phi_{local}$

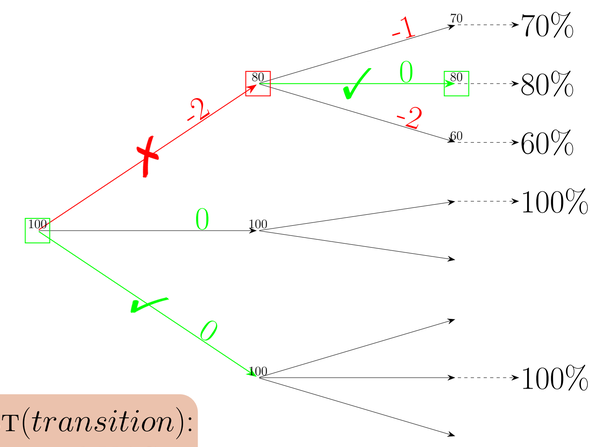
Early update
Max-violation

[Collins & Roark, 2004]
[Huang et al., 2012]



Dynamic oracle

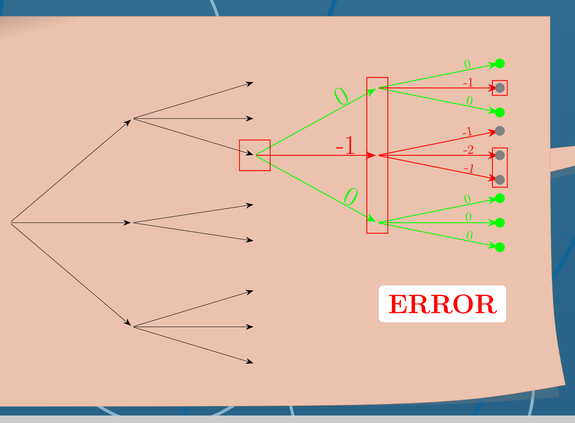
[Goldberg & Nivre, 2012]



Cost(*transition*):
Δ expected UAS

Global dynamic oracle

↪ error criterion for beam search
starting from any configuration

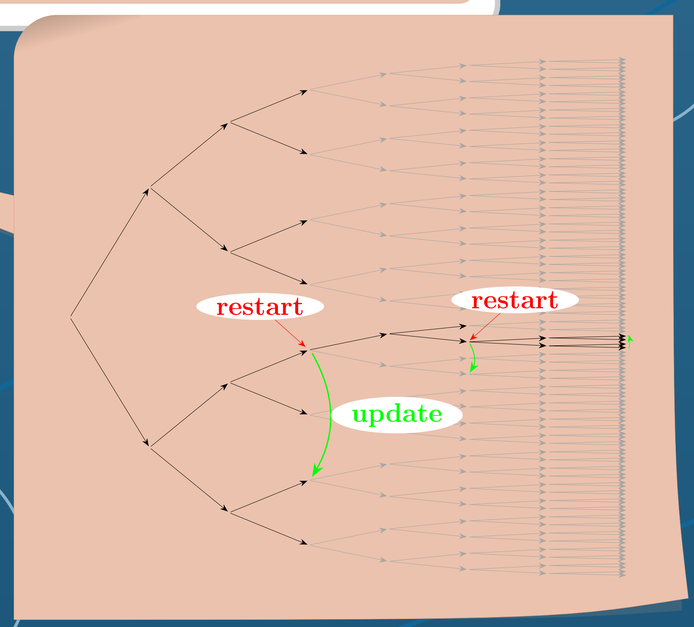


For $c' = c \circ t_1 \circ \ldots \circ t_n$:

$\text{Correct}_y(c'|c)$
$\iff \text{Cost}_y(t_1) = \cdots = \text{Cost}_y(t_n) = 0$

The reference is **never explicitly** computed

New error criterion:
when no hypothesis in beam is correct



3 sampling biases

- In case of spurious ambiguity: static choice of a **single reference**
  ↪ Inconsistent updates degrade the accuracy for current example [*typically 15% of updates*]
- Update mostly on prefix partial derivations (typical coverage: 60-80%)
  ↪ Features specific to derivation endings are **under-represented** [*punctuation marks, SOV verbs,* Root...]
- All references are sampled from the gold derivation space
  ↪ The model is not aware of the **accuracy** of non-gold parse trees

Proposal

- ▶ Use a **non-deterministic** oracle
- ▶ **Restart** on the same example after an update
- ▶ Restart with **exploration**

Improved accuracy

Evaluation on the SPMRL treebanks [9 languages]

ΔUAS	min	max	average
Early	-0.05	+0.45	+0.21
MaxV	-0.02	+0.70	+0.20

- Never hurts training
- For each language: significant gains for at least one version

Specific improvements on the **latter part** of the sentence [French: 86.02 → 86.26]

Quarter	1st	2nd	3rd	4th
Early	90.0	85.4	83.1	84.7
imp-Early	90.0	85.3	**84.2**	**85.1**

Improved sampling

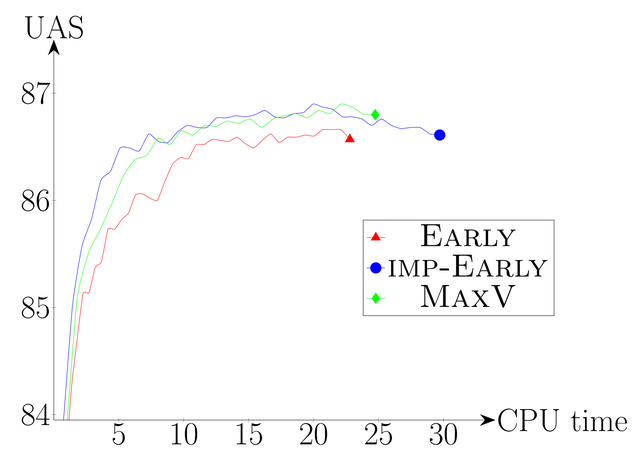
Train/test feature distributions [French]

KL divergence	Baseline	Improved
Early	0.350	0.280
MaxV	0.357	0.277

Training configurations better resemble **the actual prediction task**: configurations seen at test time

Improved convergence

Learning curves on the validation set [French]



EACL'17 – European Chapter of the Association for Computational Linguistics