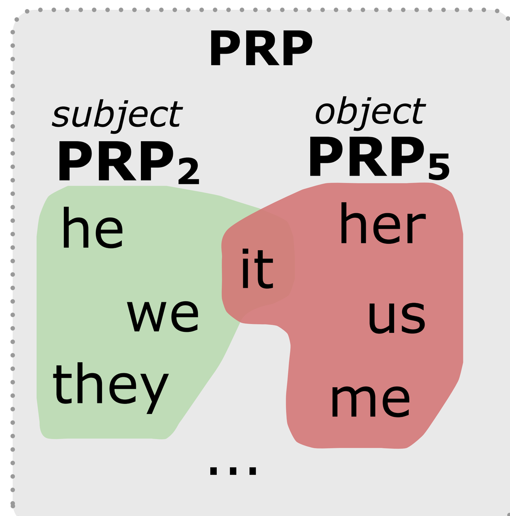
human language technology
center of excellence

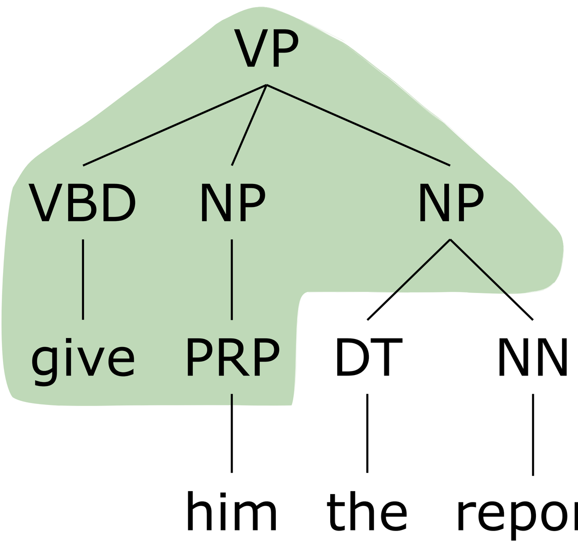
Toward Tree Substitution Grammars with Latent Annotations

Francis Ferraro, Benjamin Van Durme and Matt Post • Johns Hopkins University

Motivation

PCFGs may be overly permissive and make unrealistic independence assumptions. Automatically-learned **latent annotations** of PCFG symbols can help address this issue, and work well at the lexical level (Matsuzaki et al., 2005; Petrov et al., 2006).





Tree substitution grammars (TSGs) have an extended domain of locality and can capture long-range grammatical dependencies:

*They **give him the report**.*

➤ These approaches are complementary: can we learn latently annotated TSGs?

Automatically-Induced Latent PCFGs

Matsuzaki et al., (2005) split all categories equally and learned weights via EM. Petrov et al., (2006) used an iterative split-merge EM framework, which recovered many of the splits manually determined by Klein and Manning, (2003) and allowed symbols to have differing numbers of refinements.

Learning Probabilistic TSGs

DOP

Extract all subtrees from observed corpora: learned fragments can be large and not generalizable.

Bayesian Methods

A DP Prior encourages compact fragments (Cohn et al., 2009; Post et al., 2009); they are non-deterministic and can be complex to implement.

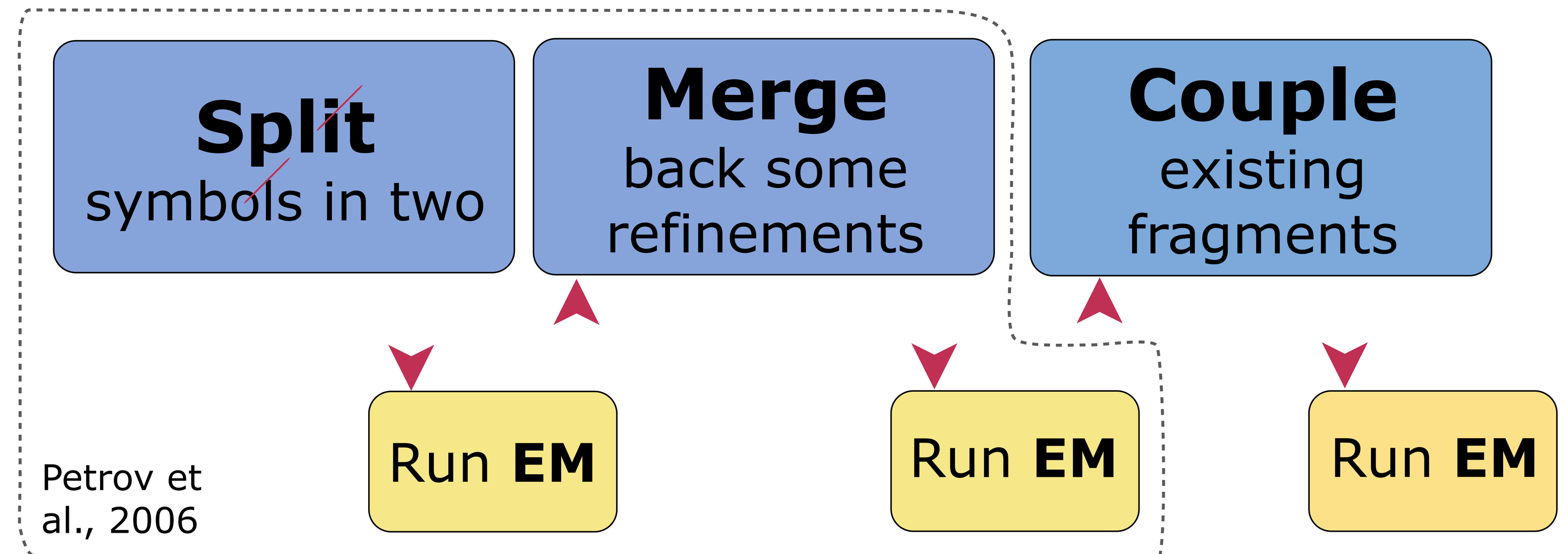
➤ Can we combine the intuitions motivating the Bayesian approach with the simplicity of DOP?

Implementation Notes

PTSG as PCFG: Make the root of every internal depth-one subtree unique and place the entirety of the TSG weight on the root depth-one rule.

Control exponential growth: (1) use binary trees; (2) forbid multiple frontier nodes from simultaneously becoming internal nodes ("chained" couplings); and (3) allow couplings only if permitted by a **constraint set** C.

Algorithm Overview



Given a grammar **G** and constraint set **C**, we **couple** by:

1. Constructing a grammar **G′** from **G** and allowed couplings from **C**

$$\mathbf{G}' = \mathbf{G} \cup \{X \circ Y \in \mathbf{C} \mid X \in \mathbf{G}\}$$

2. Estimating initial **G′** fragment weights
3. Fitting weights of **G′** via inside/outside

Deriving a Constraint Set

We deterministically count compact TSG fragments via dynamic programming: iteratively extract the *K* most frequent subtrees of size *R*. These parameters enforce sparsity and help temper exponential growth.

EXTRACTFRAGMENTS(R,K)

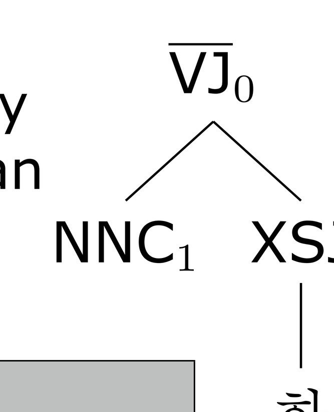
```
S ← ∅
F(1,K) ← top K CFG rules used
for r = 2 to R do
  S ← {F ∈ F(r-1, K), extended by 1 rule}
  F(r,K) ← top K elements of F(r-1, K) ∪ S
end for
```

Qualitative Evaluation

We perform a qualitative analysis of fragments learned on the Korean Treebank v2.0 and Sect. 2-3 of Penn Treebank (WSJ). We also experiment with Petrov et al. (2011)'s universal tag set, and further replace all preterminals with a single symbol, "X."

Korean Treebank

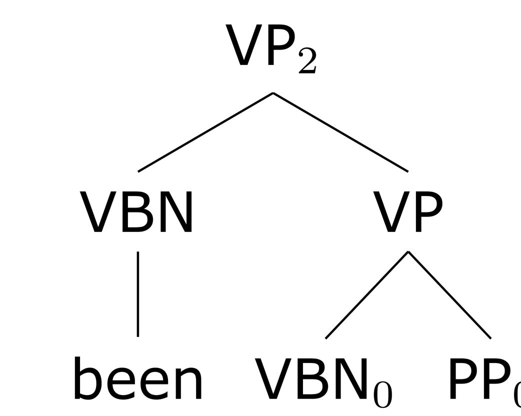
Common nouns refinements yield temporally representative nouns (NNC_0), while NNC_1 can be verbally inflected, and NNC_2 can be adjectivally inflected.

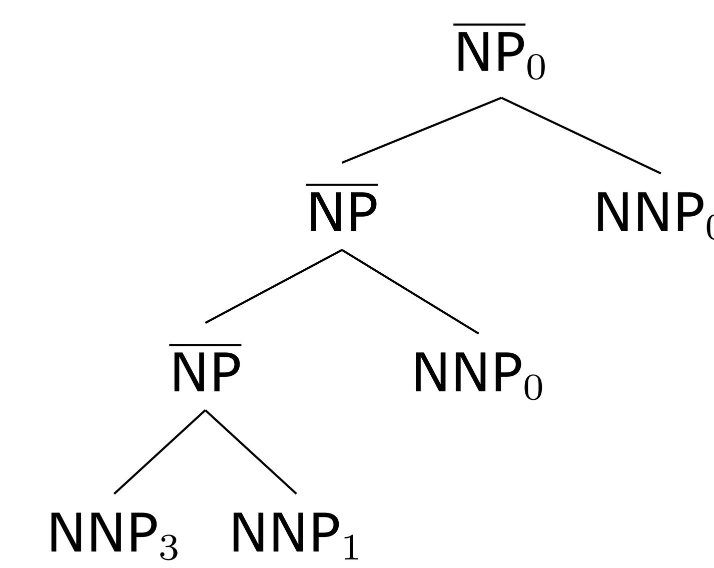


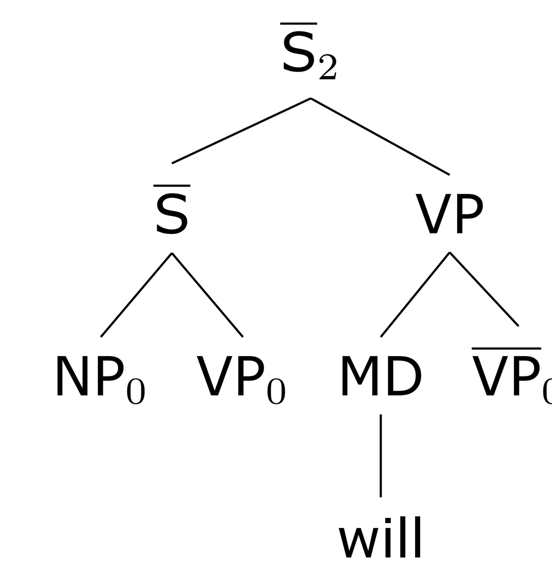
	NNC
0	경우 (*case*), 이날 (*this day*) , 현재 (*at the moment*)
1	국제 (*international*), 경제 (*economy*), 세계 (*world*)
2	관련 (*relation*), 발표 (*announcement*), 보도 (*report*)

WSJ (Sect. 2-3)

We learn descriptive lexicalized and unlexicalized fragments. We learn perfective constructions (→), a four-step modal construction (↘), and potentially useful extended nominals (↓).



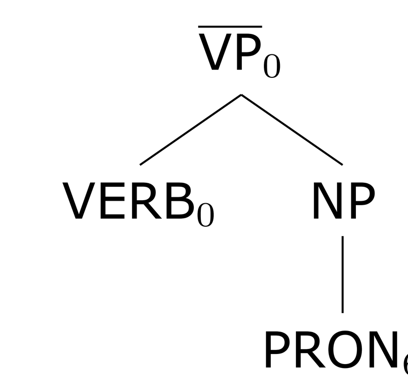




Universal Tag Set, WSJ (Sect. 2-3)

Using a coarse "universal" part-of-speech tag set we learn lexical clusters. Linguistic constraints are respected in TSG fragments, e.g., correctly placing $PRON_6$ in accusative position.

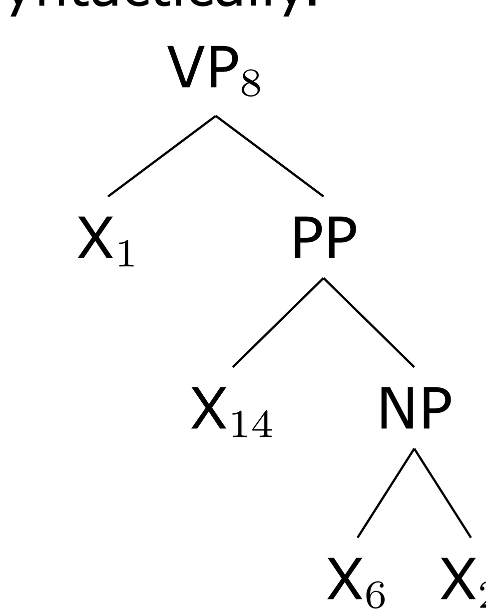
	PRON
1	its, his, your
3	what, whose, What
5	it, he, they
6	it, them, him



Preterminals as "X," WSJ (Sect. 2-3)

Recovery of the universal tag set is promising: refinements reasonably correspond with open- and closed-class distinctions, which interact syntactically.

	X	Universal Tag
0	two, market, brain	NOUN
2	%, company, year	NOUN
1	's, said, yes	VERB
13	is, was, are	VERB
3	it, he, they	PRON
12	which, that, who	PRON
6	the, a, The	DET



References and Acknowledgements

T. Cohn, S. Goldwater, and P. Blunsom. Inducing compact but accurate tree-substitution grammars. NAACL 2009.
T. Matsuzaki, Y. Miyao, and J. Tsujii. Probabilistic CFG with latent annotations. ACL 2005.
S. Petrov, L. Barrett, R. Thibaux, and D. Klein. Learning accurate, compact, and interpretable tree annotation. ACL 2006.
S. Petrov, D. Das, and R. McDonald. A universal part-of-speech tagset. *ArXiv*, April 2011.
M. Post and D. Gildea. Bayesian learning of a tree substitution grammar. ACL 2009.

We would like to thanks Byung Gyu Ahn for graciously helping us analyze the Korean results.