# Co-Training methods for Statistical Parsing using Lexicalized Grammars

Anoop Sarkar
Dept. of Computer and Information Sciences
University of Pennsylvania
anoop@linc.cis.upenn.edu

Open Issues in Lexicalized, Corpus-based Language Processing

- Adapting to new domains: training on one domain, testing (using) on another.
- Higher performance when using limited amounts of annotated data.
- Separating structural (robust) aspects of the problem from lexical (sparse) ones.

## Statistical Parsing: Supervised vs. Unsupervised Methods

- "Stone soup" approaches to unsupervised learning of parsers cannot handle structurally rich parses found in the Penn Treebank.
  (Lafferty et al 1992; Della Pietra et al 1994; de Marcken 1995)

- A feasible technique: Combining Labeled and Unlabeled Data
  - Active Learning: Bet on which examples are the hardest.
    (and annotate them)
  - Co-Training: Bet on which examples can be handled with high confidence. (use as labeled data)

## Case Study in Unsupervised Methods: POS Tagging

- POS Tagging: finding categories for words
- *. . . the stocks rose/V . . .* vs. *. . . a rose/N bouquet . . .*
- Tag dictionary: *rose: N, V*
  and nothing else

## Case Study: Unsupervised POS Tagging

- (Cutting et al. 1992) The Xerox Tagger: used HMMs with hand-built tag dictionaries. High performance: 96% on Brown

- (Merialdo 1994; Elworthy 1994) used varying amounts of labeled data as seed information for training HMMs.

  Conclusion: HMMs do not effectively combine labeled and unlabeled data

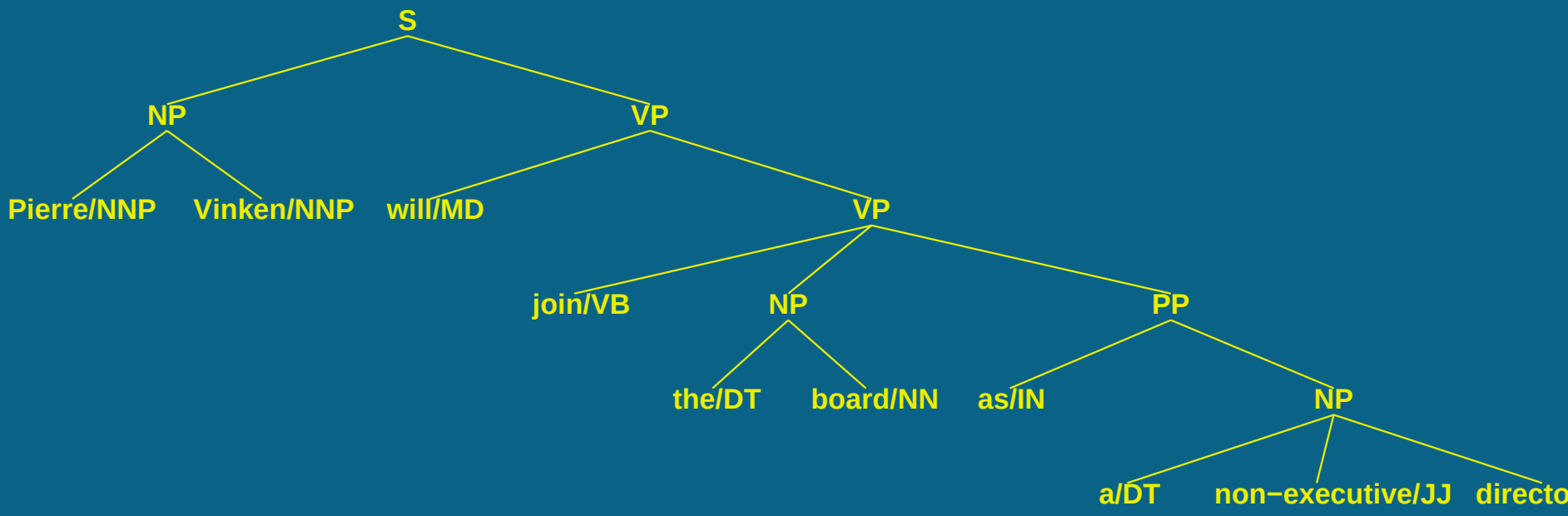
- (Brill 1997) aggressively used tag dictionaries taken from labeled data to train an unsupervised POS tagger.

  Performance: 95% on WSJ. Approach does not easily extend to parsing: no notion of tag dictionary.

## Co-Training (Blum and Mitchell 1998; Yarowsky 1995)

- Pick two (or more) “views” of a classification problem.
- Build separate models for each of these “views” and train each model on a small set of labeled data.
- Sample an unlabeled data set and to find examples that the models agree upon the most. Exploit the mutual constraints between the models
- Agreement can be computed as a simple product or in a more complex fashion. (Collins and Singer 1999; Goldman and Zhou 2000)
- Bet that these examples are good as training examples and iterate.

Pierre Vinken will join the board as a non-executive director



## Recursion in Parse Trees

- Usual decomposition of parse trees:

  S(join) $\rightarrow$ NP(Vinken) VP(join)

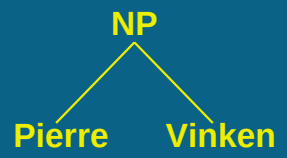
  NP(Vinken) $\rightarrow$ Pierre Vinken

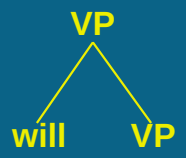
  VP(join) $\rightarrow$ will VP(join)

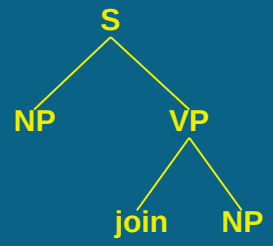
  VP(join) $\rightarrow$ join NP(board) PP(as)

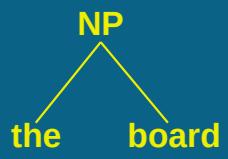
  . . .

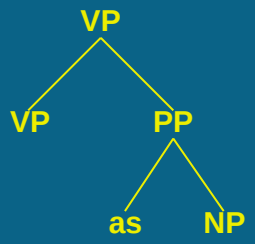
# Parsing as Tree Classification and Attachment



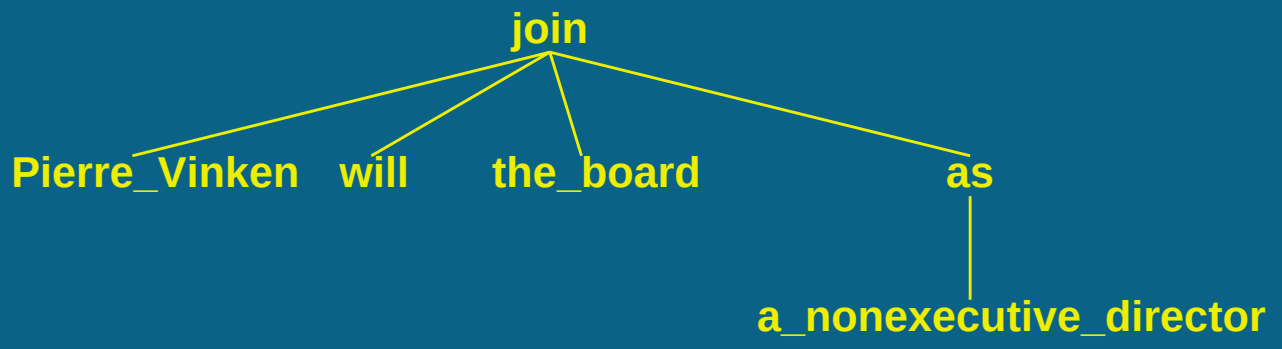





Model H1: $P(T_i \mid T_{i-2}T_{i-1}) \times P(w_i \mid T_i)$

## Parsing as Tree Classification and Attachment




Model H2: $\boldsymbol{P}(\text{TOP} = \boldsymbol{w}, \boldsymbol{T}) \times \boldsymbol{\Pi_i P}(\boldsymbol{w_i}, \boldsymbol{T_i} \mid \boldsymbol{\eta}, \boldsymbol{w}, \boldsymbol{T})$

## The Co-Training Algorithm

1. Input: *labeled* and *unlabeled*
2. Update cache
   - If *unlabeled* is empty; exit
   - Randomly select sentences from *unlabeled* and refill *cache*
3. Train models H1 and H2 using *labeled*
4. Apply H1 and H2 to cache: compute $P(T_0, \ldots, T_N)$
5. Multiply values for $P(T_0, \ldots, T_N)$ from H1 and H2 and renormalize
6. Pick most probable $n$ from renormalized score
7. Remove best $n$ from *cache* and add to *labeled*
8. $n = 2n$; Go to Step 2

## Preliminary Experiment

- *labeled* was set to Sections 02-06 of the Penn Treebank WSJ (9625 sentences)
- *unlabeled* was 30137 sentences (Section 07-21 of the Treebank stripped of all annotations).
- A TAG dictionary of all lexicalized trees from *labeled* and *unlabeled*.
  Novel trees were treated as unknown tree tokens
- The *cache* size was 3000 sentences.

## Preliminary Experiment

- Test set: Section 0 (*development test set*)
- Baseline Model was trained only on the *labeled* set:
  Labeled Bracketing Precision = 67.43% Recall = 64.93%
- After 12 iterations of Co-Training:
  Labeled Bracketing Precision = 81.2% Recall = 78.94%
- NEW!: Evaluation of an unsupervised approach is directly comparable to other supervised parsers.

## Summary

- Methods that combine labeled and unlabeled data provide a promising new direction towards unsupervised learning.
- Co-Training, previously used for classifiers with 2/3 labels, was extended to the complex problem of statistical parsing.
- Parsing treated as providing structured (tree) labels with attachments computed between these labels.
- Evaluation of a unsupervised method for parsing directly comparable with supervised approaches.

## Future Work

- Current Work: Improve parser (better smoothing); Better combination of the models.
- Experiment with using a larger labeled (1M words) and unlabeled set (23M words).
- Use machine learning for learning the tag dictionary:
  Thesis work: subcategorization frame learning, learning verb classes
- Conjecture: Active Learning and Co-Training can be combined into a single framework.