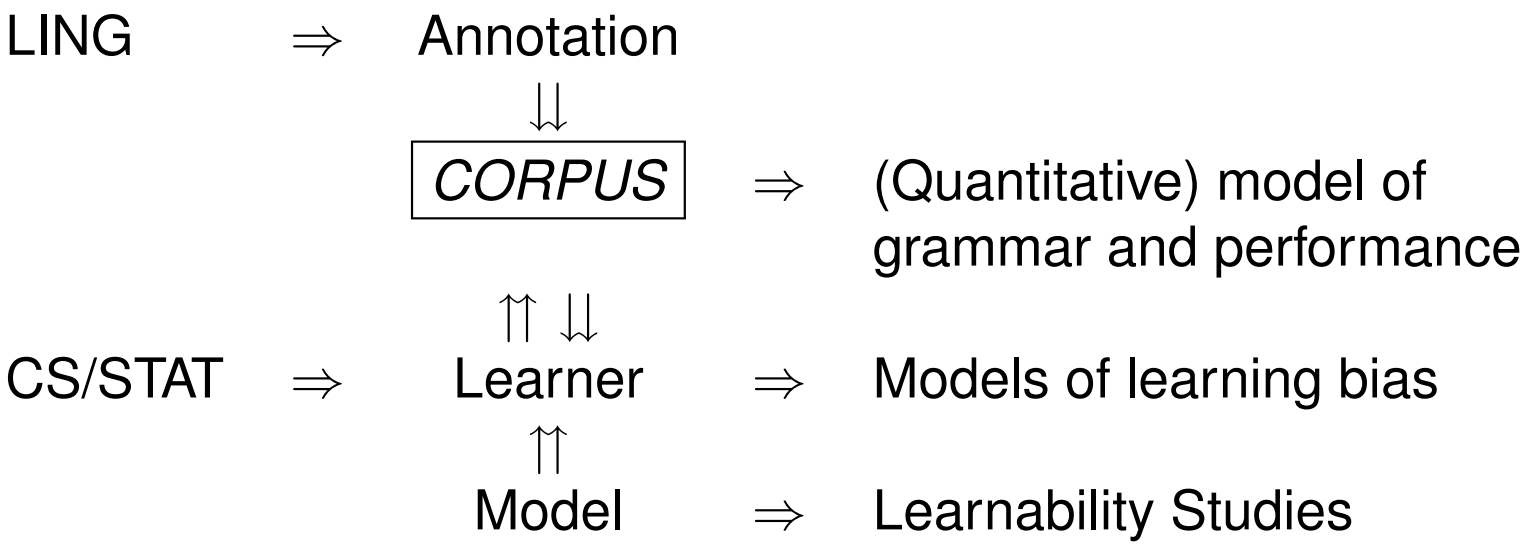
Semantic Parsing for High-Precision Semantic Role Labelling

Paola Merlo

University of Geneva
Joint work with Gabriele Musillo.
Thanks to James Henderson, Ivan Titov, and the Swiss National Science Foundation.

The Supervised Learning Methodology



Shallow Natural Language Understanding

- Dialogue
  I would like to reserve a flight from Geneva to Boston
  reserve(THEME=flight, SRC= Geneva, DIR=Boston)
- Machine Translation
  I like it = EXP PRED THEME
  Mi piace = (THEME) EXP PRED

Different Methods for Semantic Role Labelling

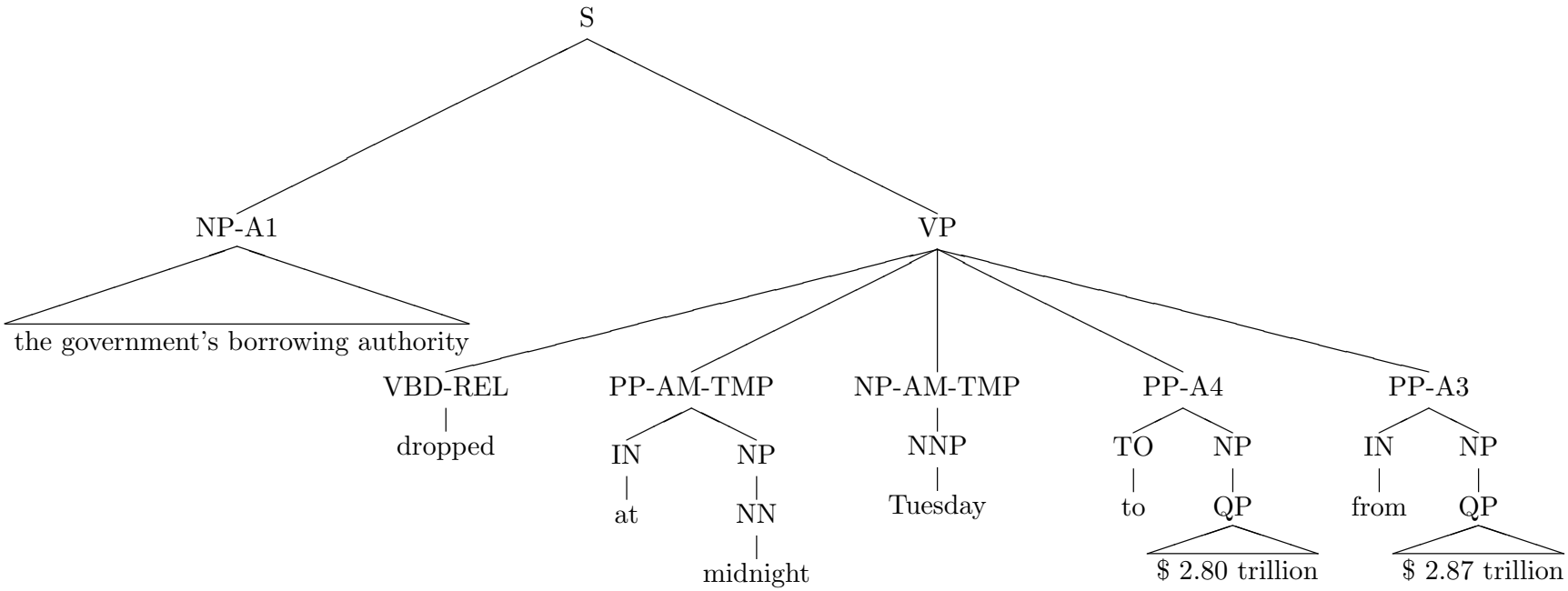
- Rule-based: manipulation of explicit knowledge representation
- Corpus-based + Machine Learning Algorithms
  - Shallow Parsing Pipeline
  - Integrated Full Parsing

Lexical Semantics Annotation: PropBank

The executives gave the chefs a standing ovation
ARG0: the executives
REL: gave
ARG2: the chefs
ARG1: a standing ovation

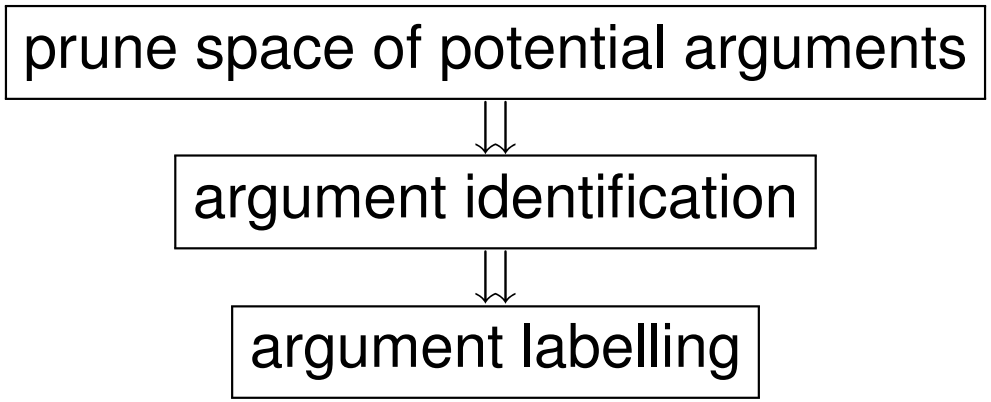
- Annotation of corpus, representative sampling
- Arguments of the predicate are annotated with **verb-specific** abstract semantic role labels A0-A5 or AA.
- Adjuncts of the predicates are annotated with **general** abstract semantic role labels that are inherited from function labels, such as AM-LOC, AM-TMP, AM-ADV.

A PropBank Tree



Shallow Semantic Processing

- Several current approaches to semantic role labelling perform the labelling in several stages



Lessons from Previous Work

- Collect statistical indicators of syntactic properties in large tagged corpus to create statistical summary of lexical semantic properties of verbs.
- Use these statistics to classify verbs into classes.
- Classes of verbs can be learnt at approximately 80% accuracy
- Verb Classification: **Syntax and semantics are highly correlated in corpus statistics**

Integrating Semantic Labelling into Parsing

- SEMANTIC ROLE LABELS CAN BE LEARNED AND RECOVERED WHILE PARSING
- Question 1: can it be done? Task is more difficult and information is less as only part of tree is available
- Questions 2: Can we learn semantic role labels robustly without parsing degradation? Despite increased data sparseness and variability.
- Question 3: what are the properties of an integrated approach to semantic role labelling and how can they be exploited at their best?

Developing a model of semantic parsing

- **Syntactic Parsing**: mapping a potentially unbounded string to a tree through a finite set of operations
- **Probabilistic Syntactic Parsing**: Probabilistic parsing handles ambiguities
- **Estimation of Probabilites**: probabilities depend on (potentially unbounded) previous history of derivation
- **Probabilistic Semantic Parsing**: based on linguistically justified correspondence between the syntactic tree and the semantic labels

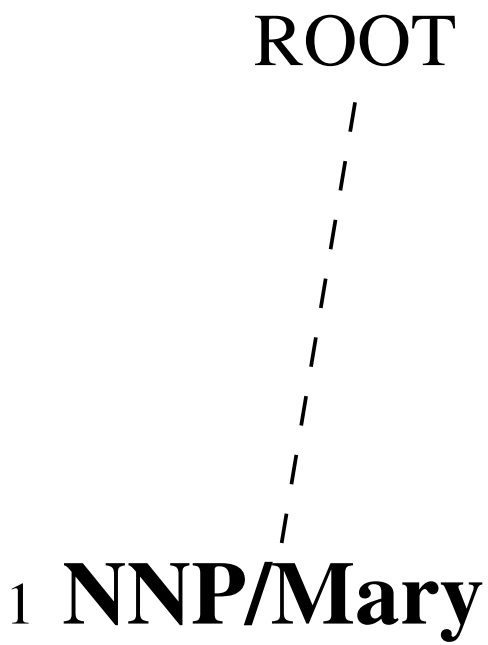
Syntactic Parsing

- **Syntactic Parsing**: mapping a potentially unbounded string to a tree through a finite set of operations
- We can represent each syntactic structure as an unbounded sequence of basic operations $d_1, d_2 \ldots d_m$, called its derivation

An Example Derivation

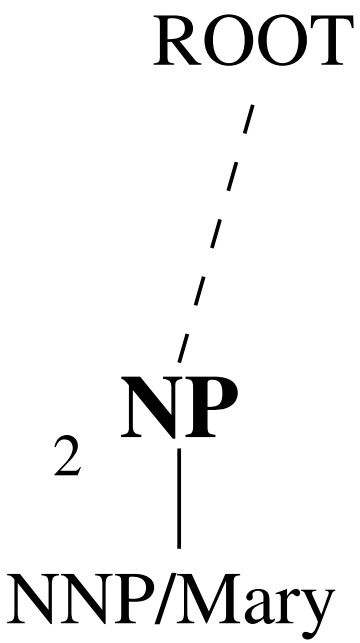
0 **ROOT**

An Example Derivation



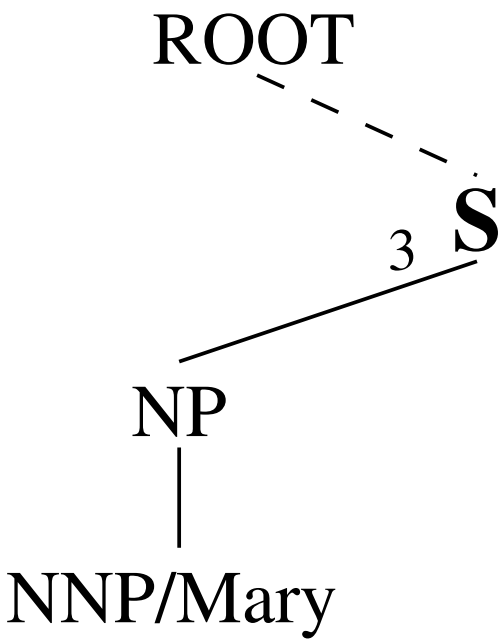
$$\mathbf{d_1 = next(NNP/Mary)}$$

An Example Derivation



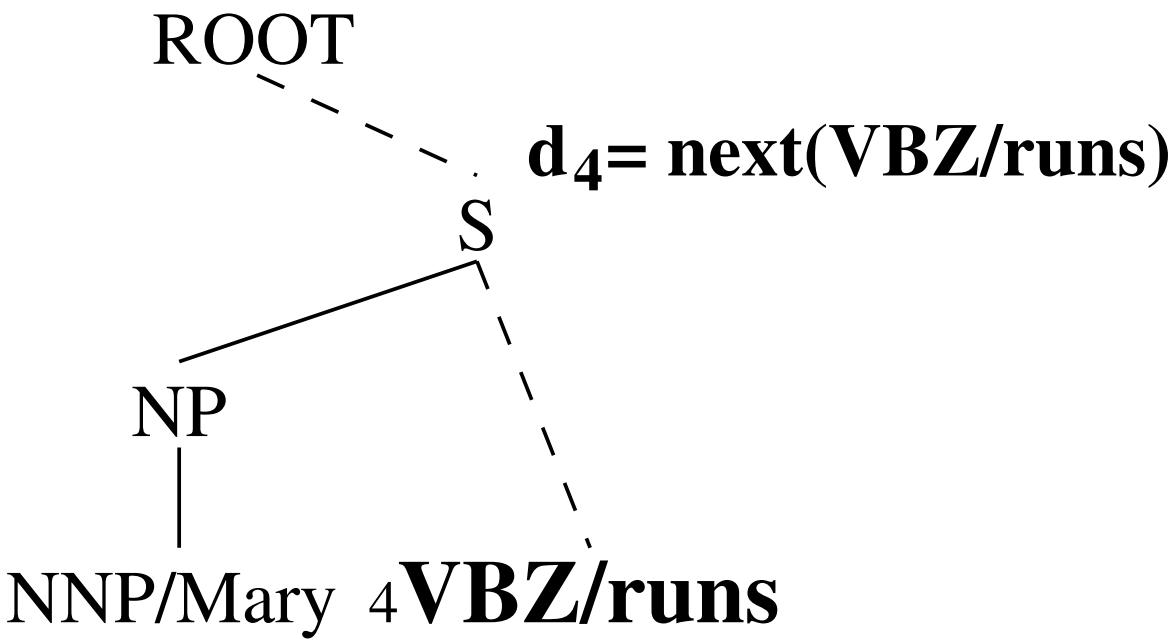
$$d_2 = \text{project}(NP)$$

An Example Derivation

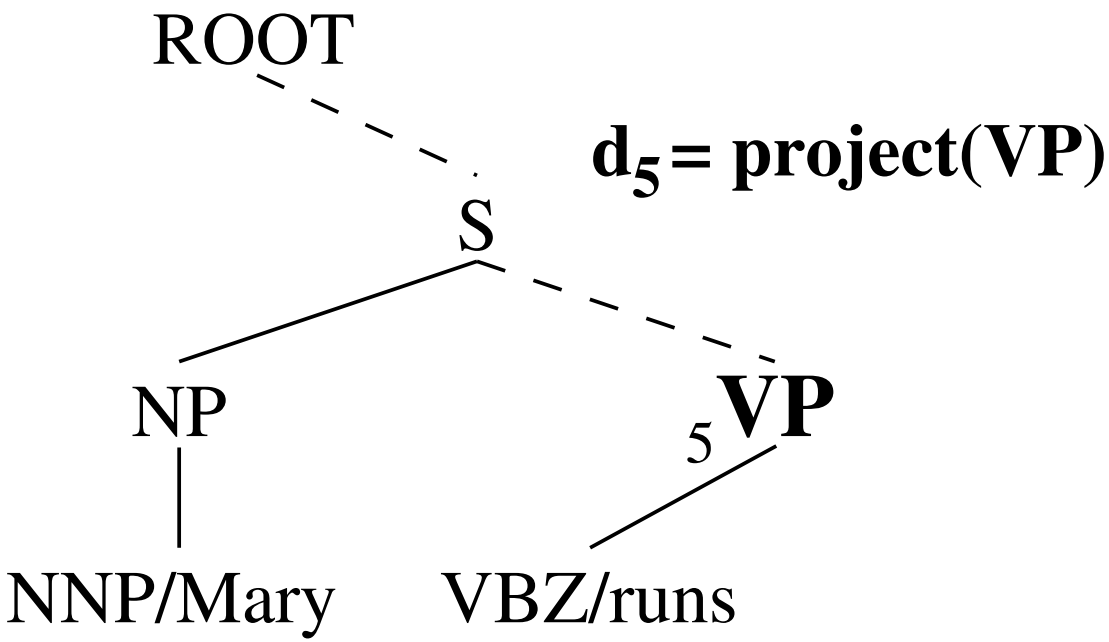


$$d_3 = \text{project}(S)$$

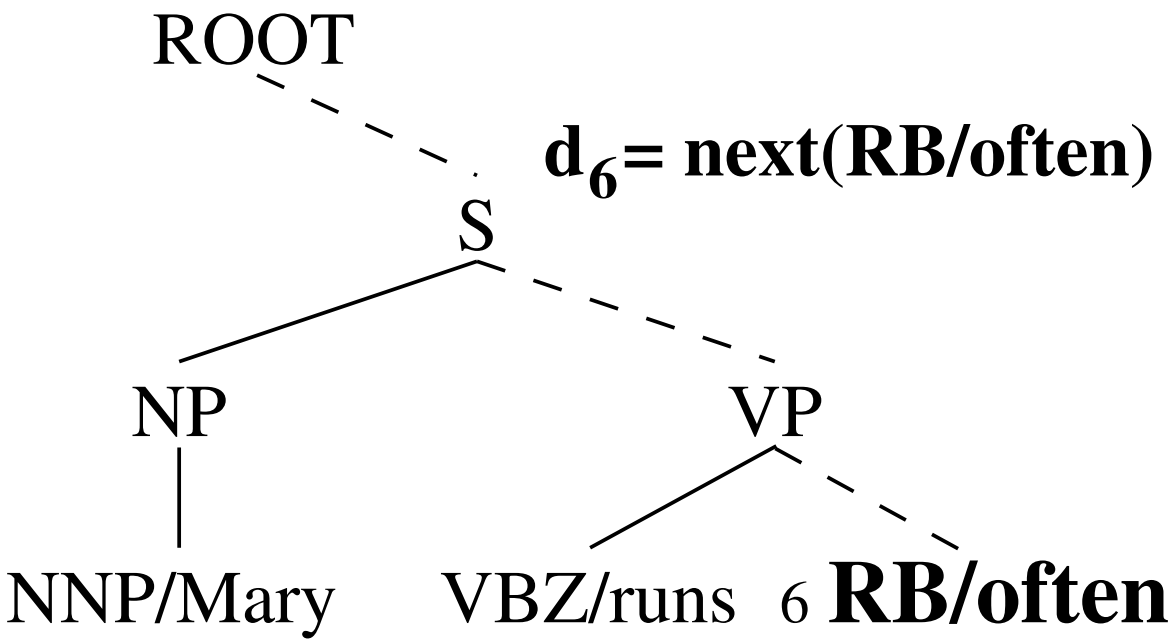
An Example Derivation



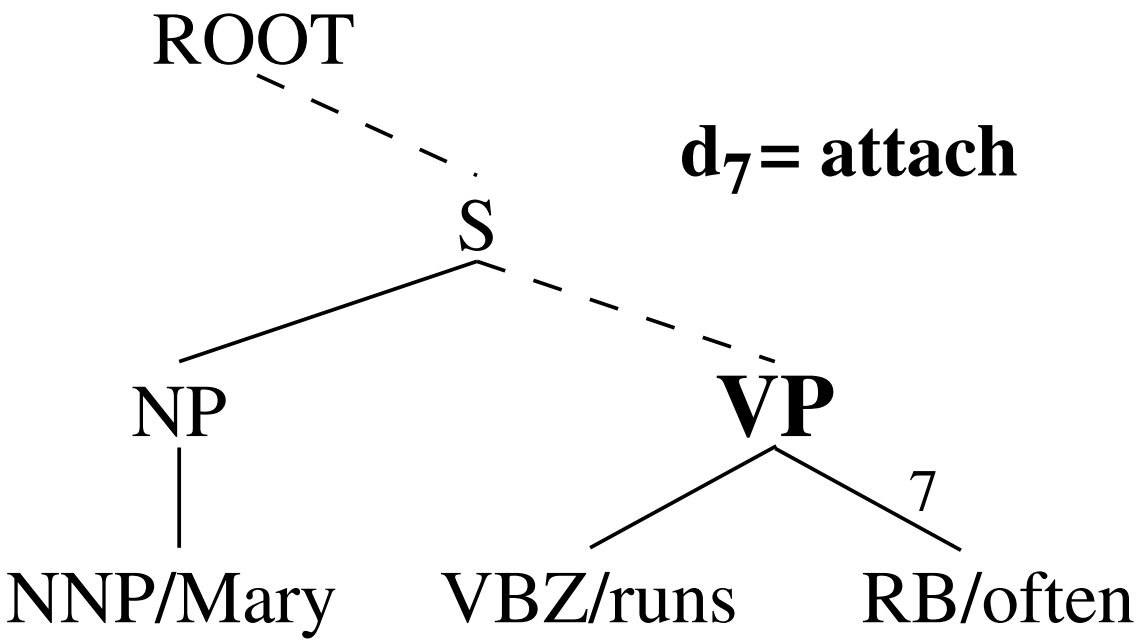
An Example Derivation



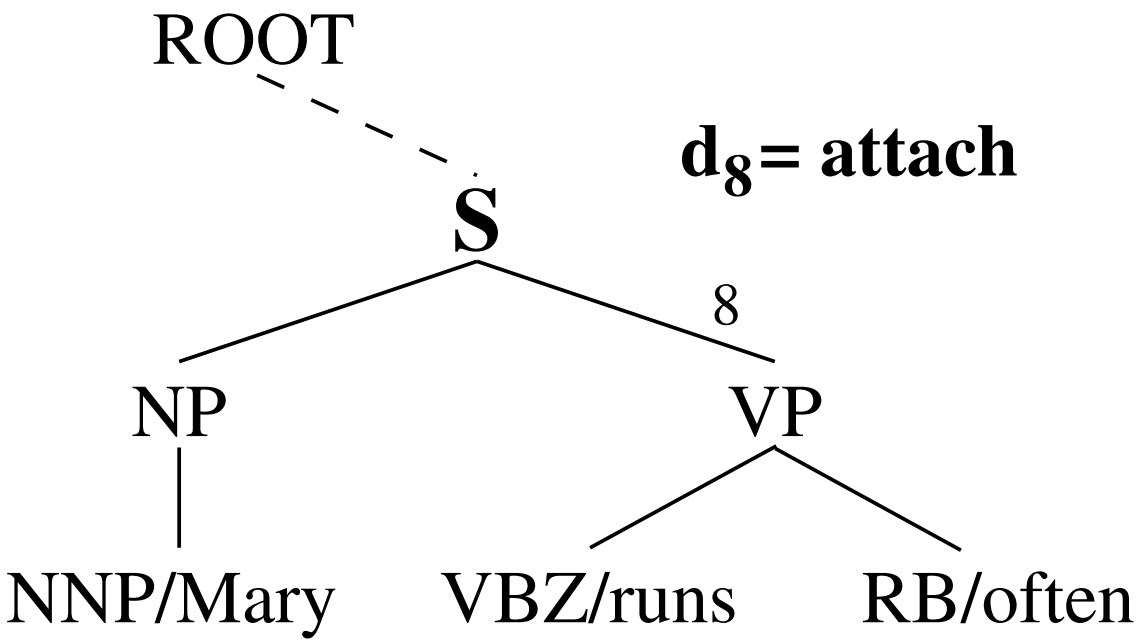
An Example Derivation



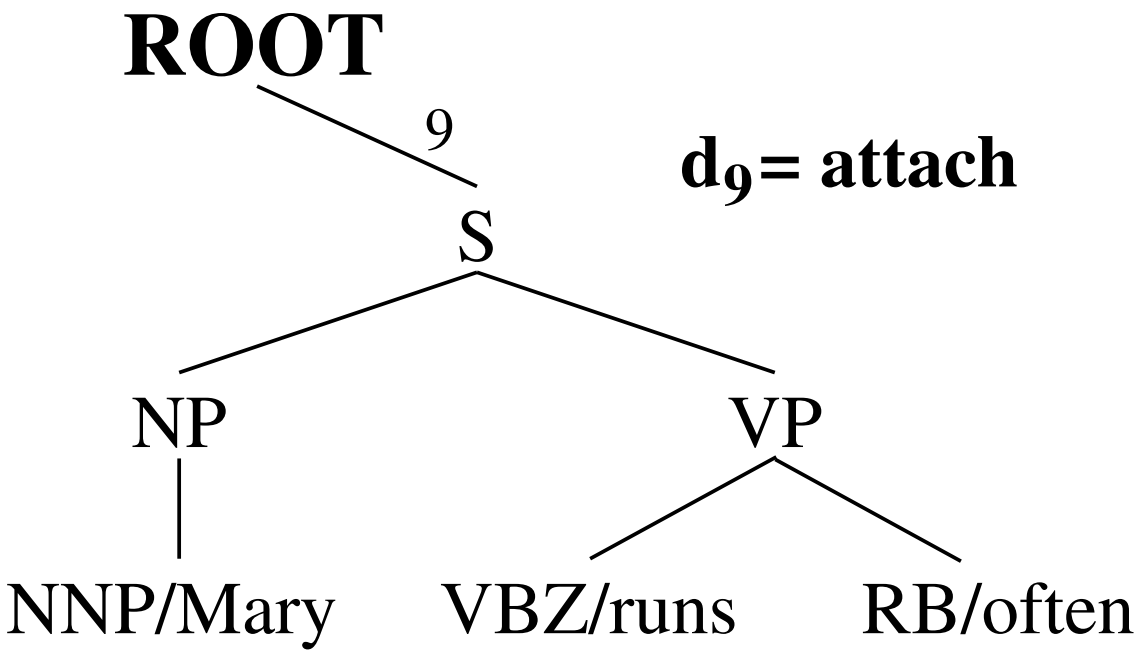
An Example Derivation



An Example Derivation



An Example Derivation



Statistical Parsing

- Some sentences are ambiguous, so sometimes they have one syntactic structure and sometime they have another
  - *I saw the man with a telescope.*
- Many sentences are too complicated for a parser to deduce a single correct structure, because the parser has incomplete information
- The best a parser can do is try to be correct as frequently as possible
- If we guess the **most probable structure**, then we will be correct the most frequently

Probability Estimation

- How do we know which syntactic structure is the most probable?

History-Based Models

- We can estimate the probability of an entire derivation by estimating the probability of each operation conditioned on its history of previous operations

$$\begin{aligned} P(d_1, d_2 \ldots d_m) &= P(d_1)P(d_2|d_1)\ldots P(d_m|d_1, d_2 \ldots d_{m-1}) \\ &= \prod_i P(d_i|d_1, d_2 \ldots d_{i-1}) \end{aligned}$$

- Now each probability only has a finite number of alternative operations

Independence Assumptions

- But these probabilities are still complicated, because the histories can be unbounded in length
- The standard approach is to make **independence assumptions**, meaning we ignore the less important information in the history
- For example, the context free assumption in Context-Free Grammars

A Neural Network Statistical Parser

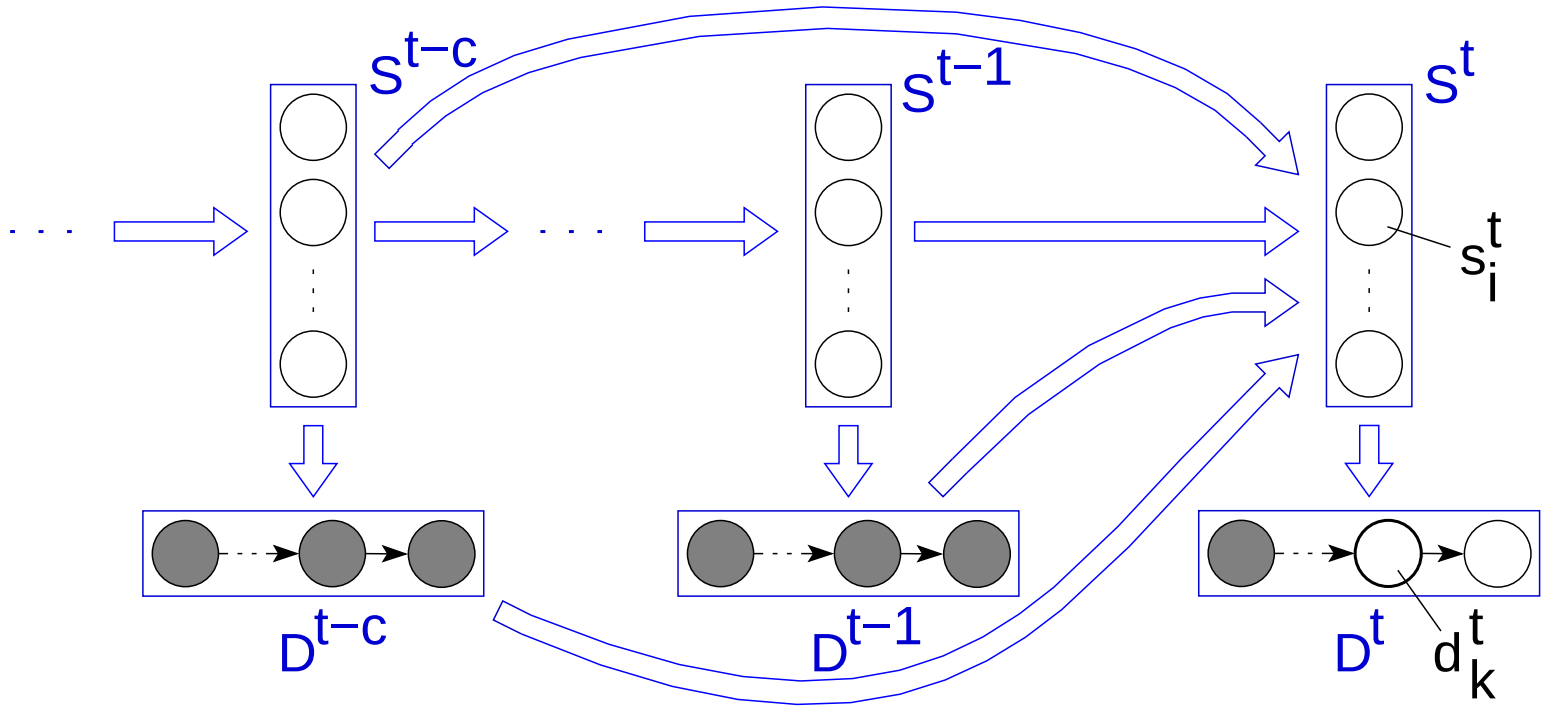
How can we estimate the probabilities $P(d_i|d_1, d_2 \ldots d_{i-1})$ without making independence assumptions (that is without simplifications that might be incorret)?

Estimating the Probabilities

- A **Simple Synchrony Network** (Henderson 2003) is trained to estimate the probabilities $P(d_i|d_1 \ldots d_{i-1})$
- The SSN first computes a **hidden** representation of the derivation history $h_i = f(d_1 \ldots d_{i-1})$ which compresses the history representation to a finite set of representational units
- The SSN then computes an **output** probability for each possible operation d_i from the hidden representation h_i

$$P(d_i|d_1 \ldots d_{i-1}) \approx P(d_i|h_i)$$

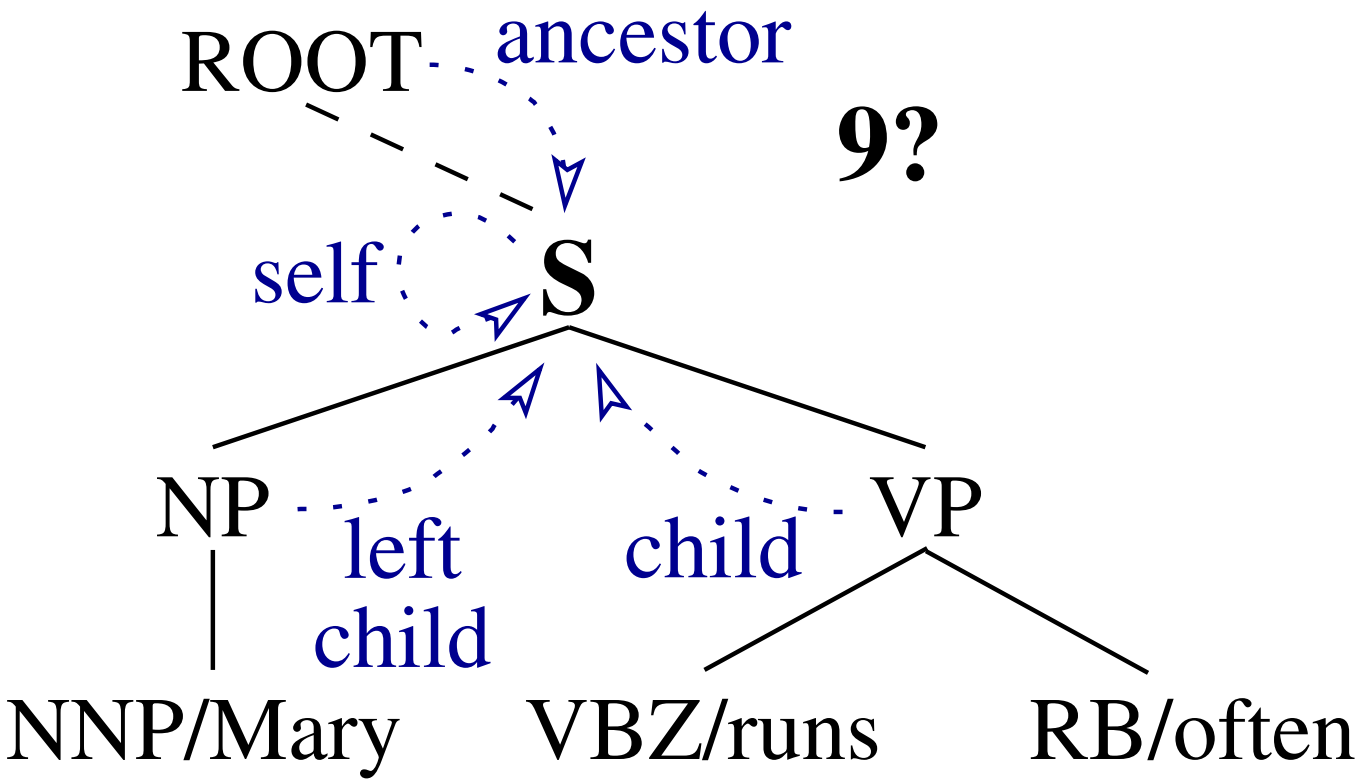
SSN Computation



Learning History Representations

- Any information about $d_1 \ldots d_{i-1}$ input at any earlier step *could* **flow** from history representation to history representation and reach the output for d_i
- Training is **biased** toward paying more attention to information which has been **input more recently** in this flow of information
- We provide an appropriate bias by matching recency in the flow of information between history representations to **structural locality** between constituents

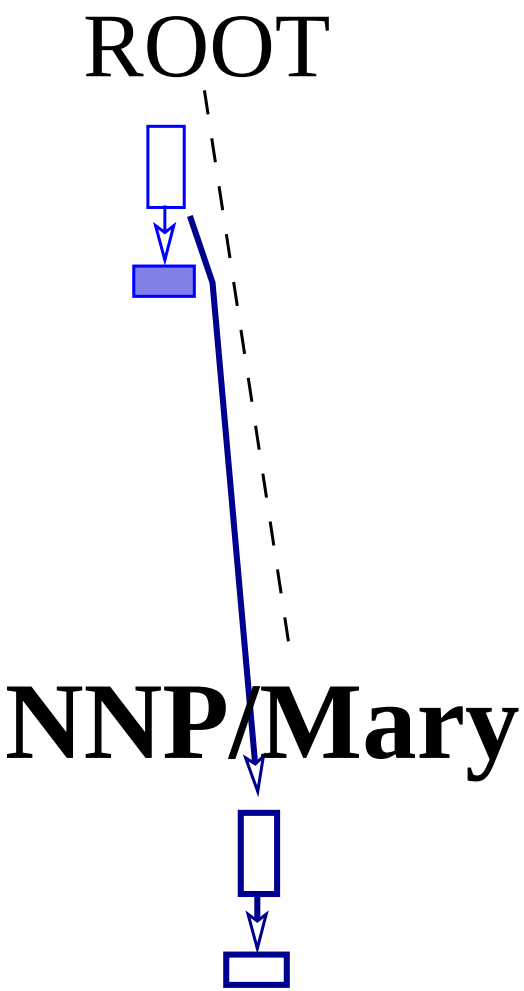
The Definition of Structurally Local



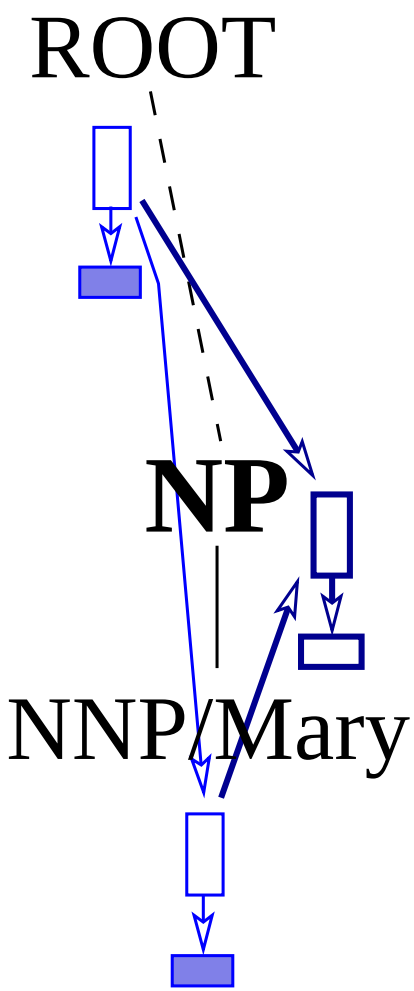
ISBN Parse Example

ROOT

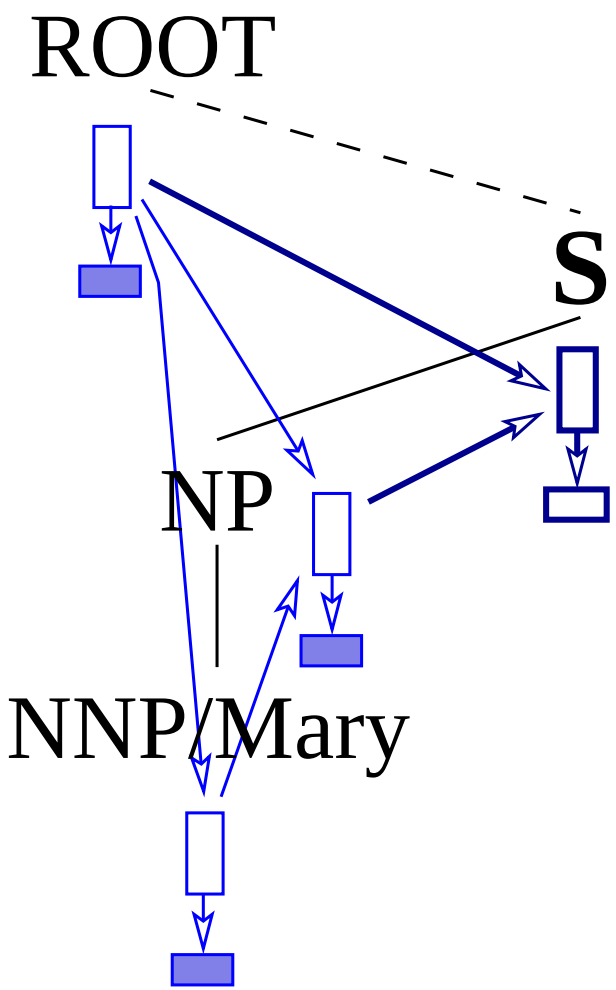
ISBN Parse Example



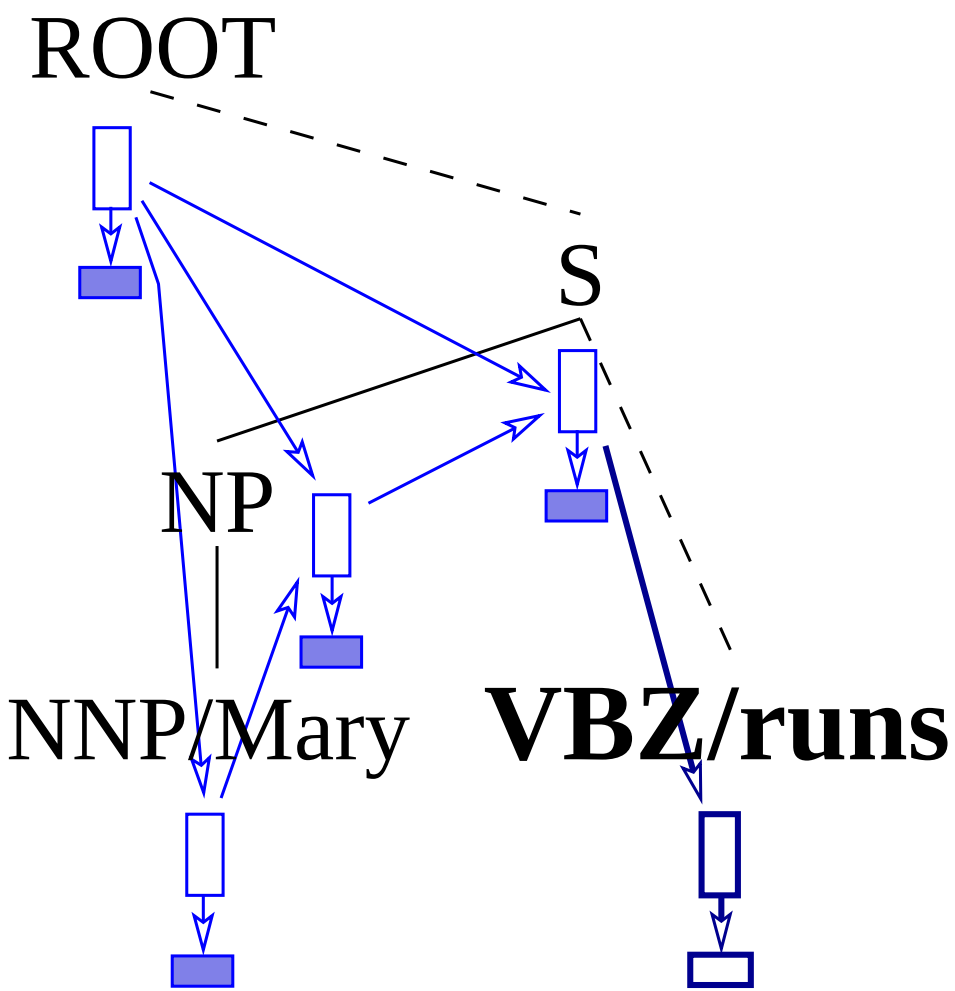
ISBN Parse Example



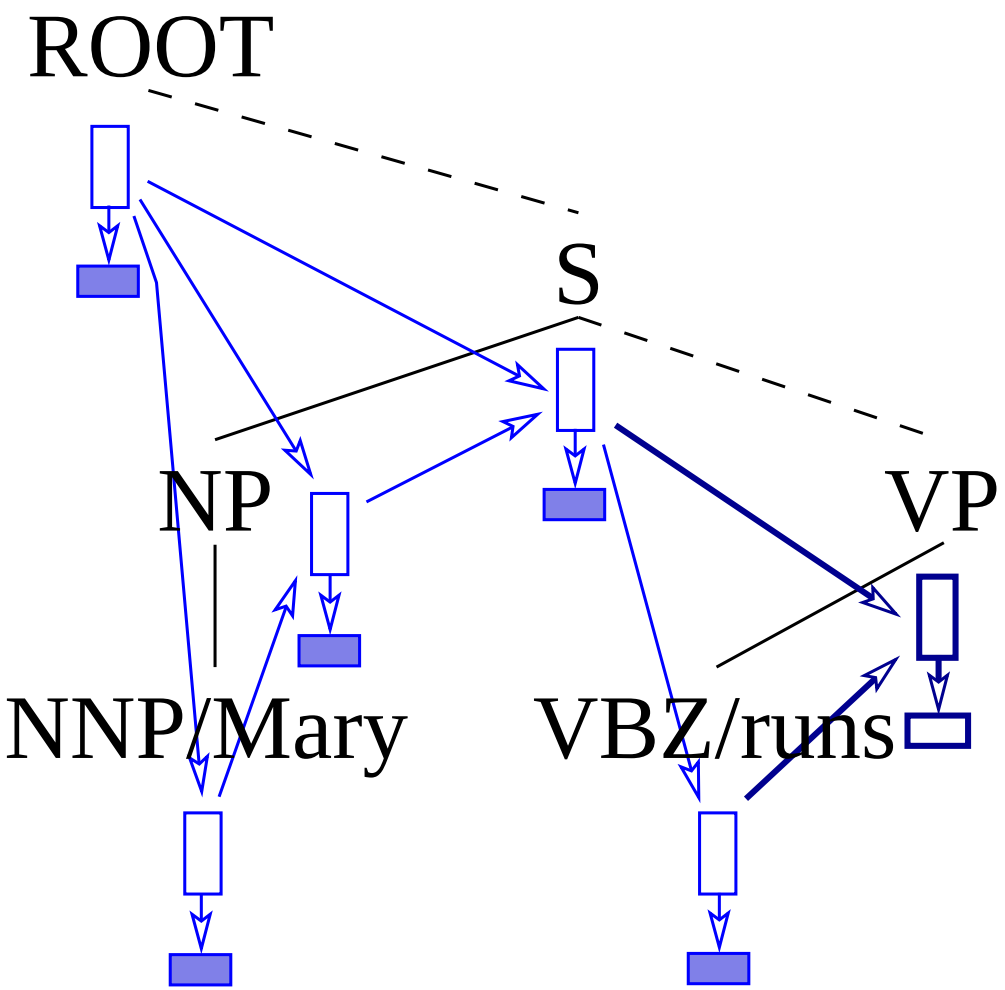
ISBN Parse Example



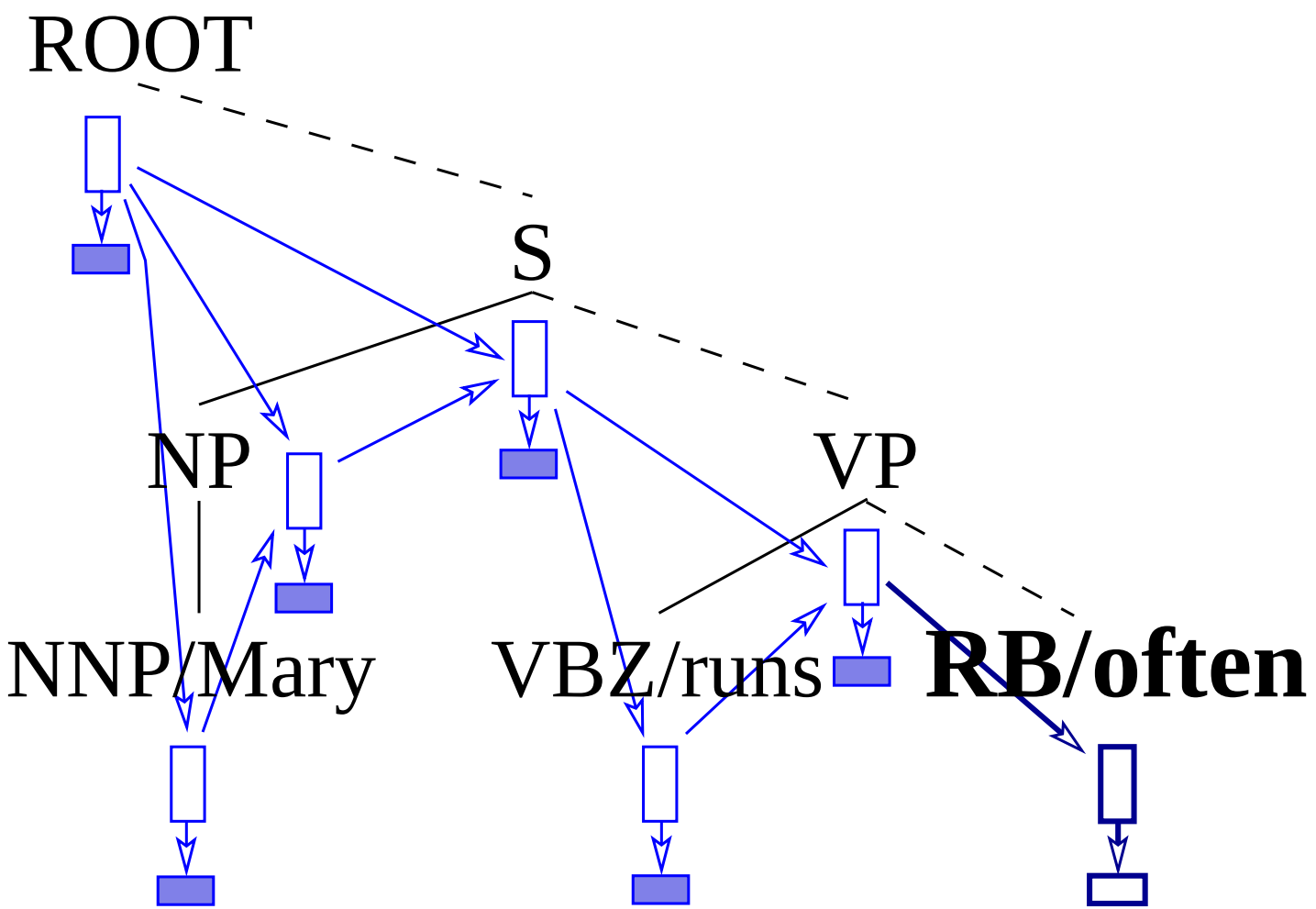
ISBN Parse Example



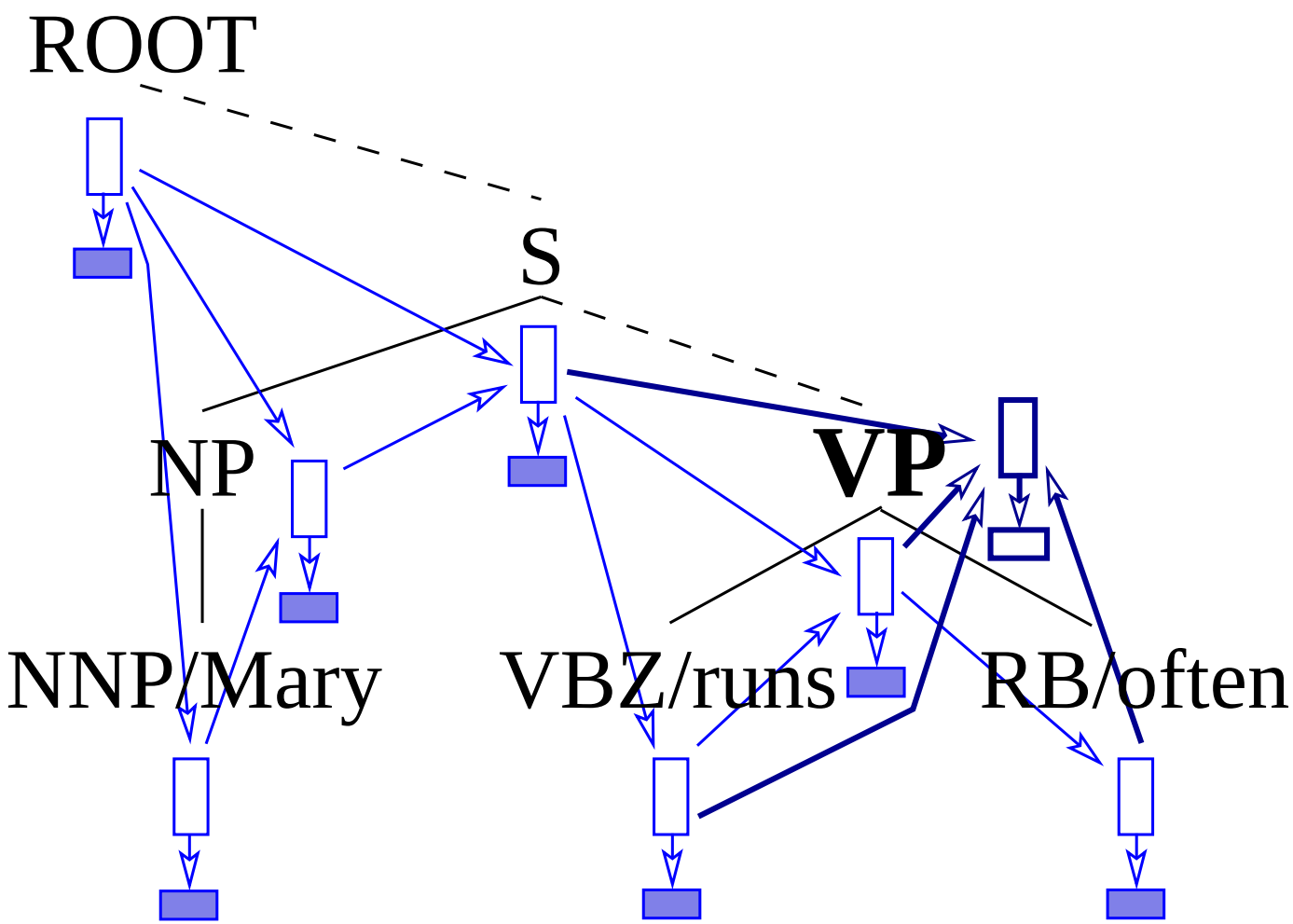
ISBN Parse Example



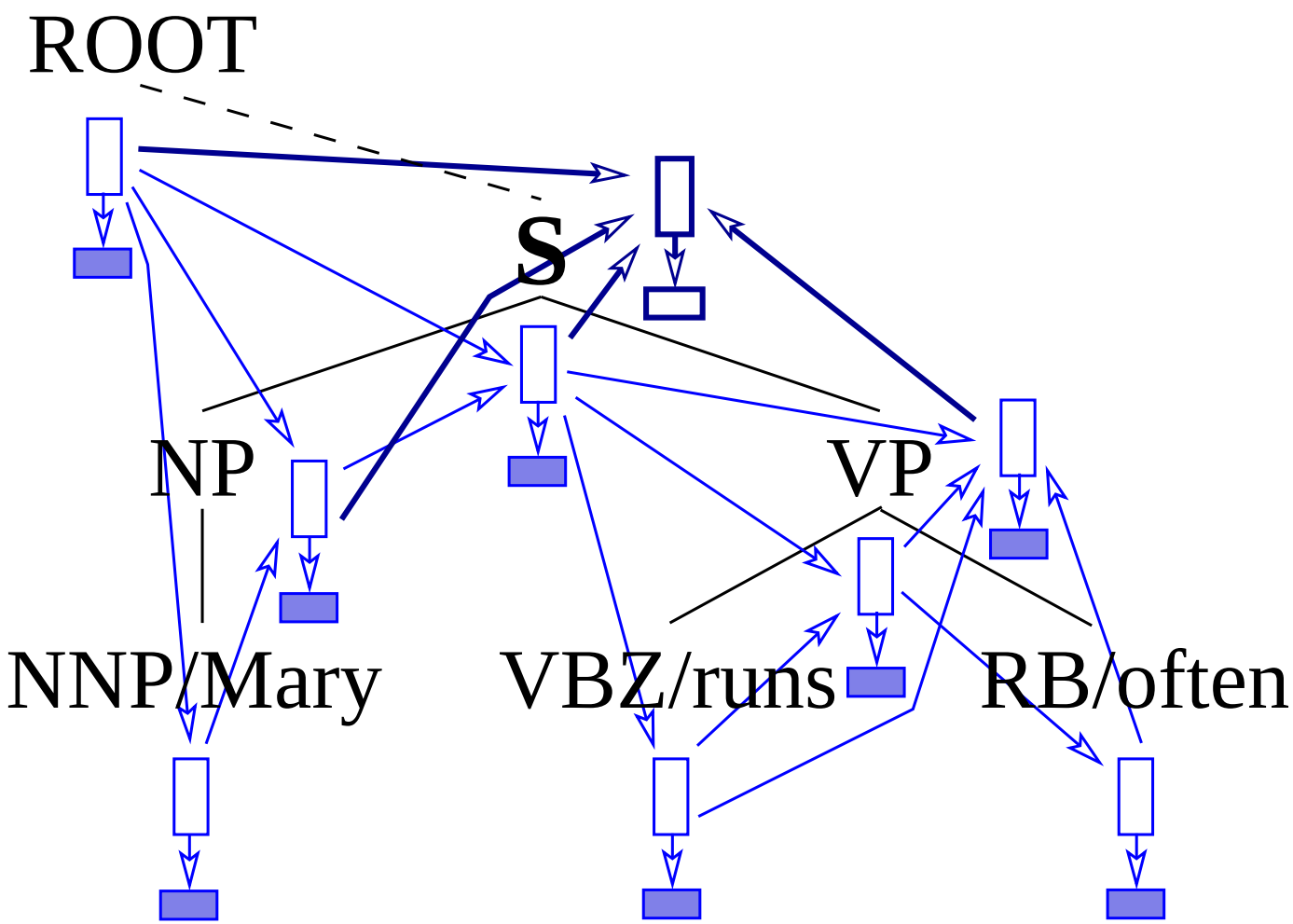
ISBN Parse Example



ISBN Parse Example



ISBN Parse Example



Extension to joint calculation of semantic labels

If we want to model semantic labels, we need to modify the parser's structural locality bias.
We introduce two kinds of biases to force the parser to pay attention to semantic role labels.

Modelling of Semantic Role Labels

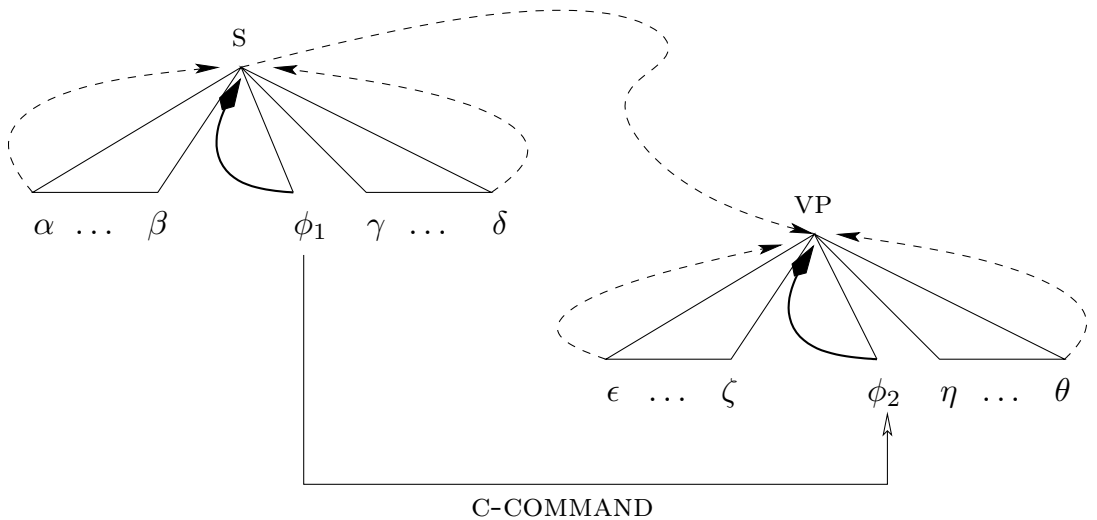
- Lexical biases: Fine-grained modelling of semantic role labelling.
- Network connectivity: highlight portion of tree that bears semantic role labels.

Changing Models' Connectivity

- Many successful semantic role labelling systems have tried to model sequences of SRLs.
- Semantic roles are mapped onto syntactic structure based on the Thematic Hierarchy
(AG > GOAL/EXP > THEME > DIR/LOC/MNR)

Structural Biases

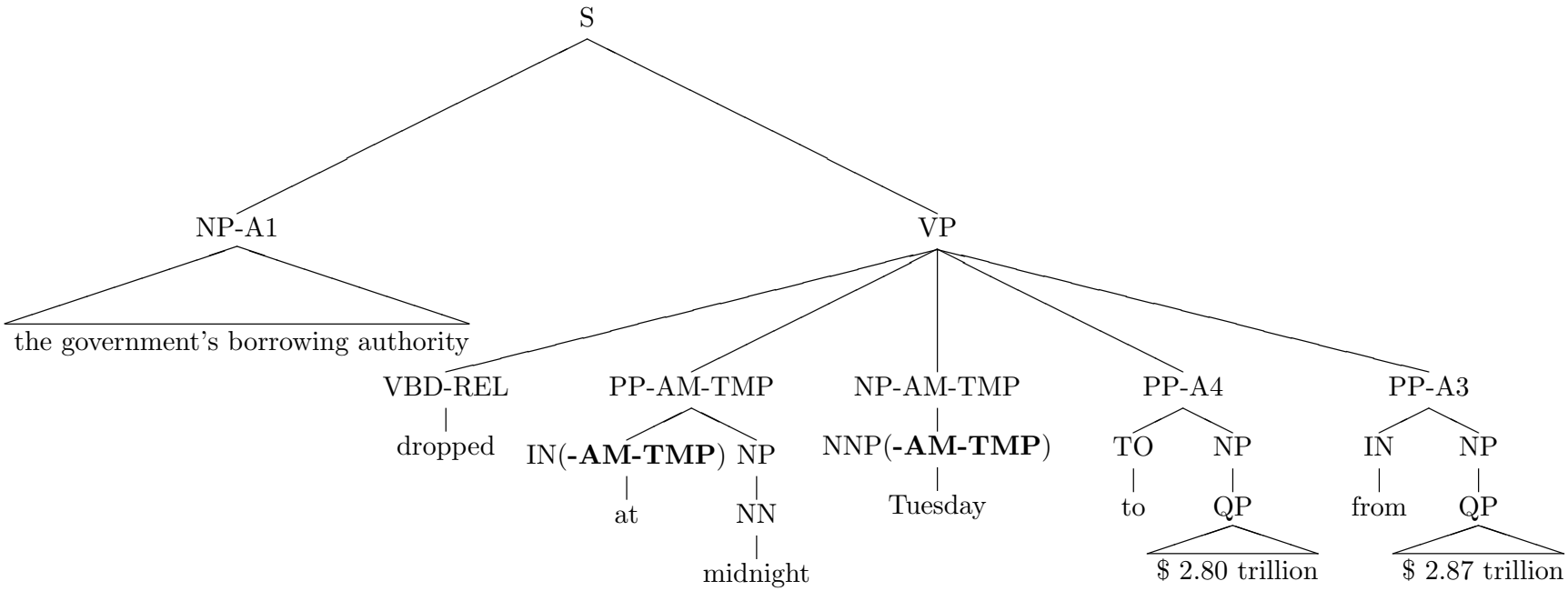
Regularities, if any, between nodes annotated with semantic labels are captured.



Introducing Lexical Biases

- Hypothesis: Semantic role labels are low in tree.
- Klein and Manning 03 have shown that a lot can be achieved by splitting tags.
- Use SR labels to split ambiguous POS tags.
- Lower SR labels (DIR,EXT,LOC,MNR,PNC,CAUS,TMP) onto tags of constituent's head.

Lowering Semantic Role Labels



The Experimental Methodology

- Standard setting for PTB learning: Training on section 2-21, testing on section 23. Cross-validation on section 24.
- SSN model has 613 labels instead of original 33. Vocabulary 4970 tag-word pairs instead of original SSN 512 pairs.
- 240 POS instead of original 45 because of label splitting.
- The baseline: do not modify parser or data in any way to establish level at which task can be performed, if at all.

Semantic Parsing Results

	Results on Development Set		
	P	R	F
Baseline	79.6	78.6	79.1
Split tags	80.5	79.4	79.9
Split tags + enhanced connectivity	81.6	80.3	81.0

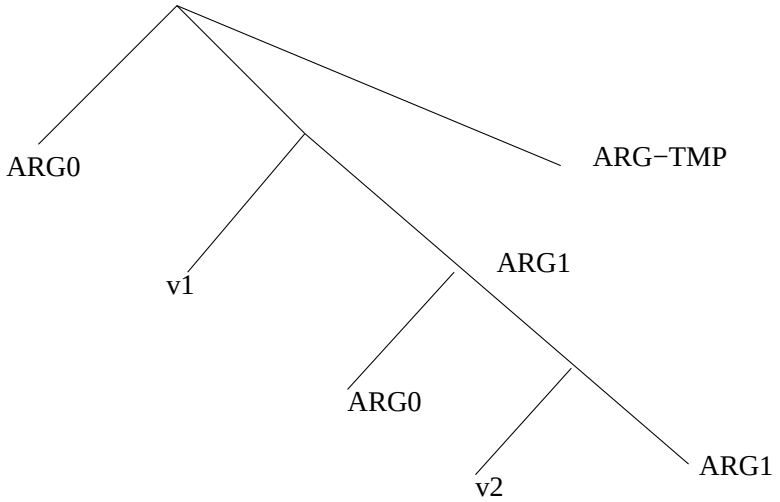
Indicative Comparison to Pipeline Architectures

	Results on Test Set		
	F	R	P
Sem-syn SSN	81.6	81.3	81.9
Best 5 CoNLL	82.7–83.4	82.5–83.1	83.0–83.7

Discussion

- Both enhancements improve over baseline
- Semantic parsing performed almost as well as other pipeline methods which do not produce a rich output
- Syntactic parsing degraded, but still at level of some of the best parsers (88.4% F- measure compared to 88.2% Collins 99)

Application to Semantic Role Labelling



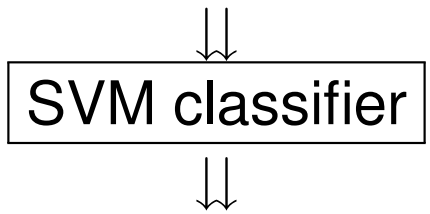
- $\Rightarrow$ (v1, ARG0, ARG1,ARG-TMP) (v2,ARG0,ARG1)

Rule-based Extraction Method

- Compile finite state automata gathered from sample of gold data defining path that connect SRL nodes to their predicate
- F-measure of 95.9% on gold data
- F-measure of 69.7% on parser output

SVM Extraction Method

verb	role	path	barrier	answer
(v1,	ARG0,	up/up/down,	noS,	yes)
(v1,	ARG1,	up/down,	noS,	yes)
(v1,	ARG-TMP,	up/up/down,	noS,	yes)
(v2,	ARG0,	up/up/up/up/up/down,	S,	no)
(v2,	ARG0,	up/up/down,	noS,	yes)



(v1,ARG1,up/down/down,S,?)

Results

Results on Development Set (CoNLL Shared Task	P	R	F
SVM All Features	87.4	63.2	73.6
Rule-based	72.9	66.7	69.7
SVM No loc/min Features	74.3	63.8	68.6
Baseline	57.4	53.9	55.6

- Usefulness of locality features

Discussion

- Conll shared task, test section
- Results globally not very competitive: bottom third compared to ensemble of learners (79.4–66.7, 75.1)
- Results compared to single systems that do not use any external knowledge: top third (76.4–74.3, 75.1)
- Best precision, low recall (87.6 vs 82.3, 65.8 vs 74.8)

Combination with high-recall systems

- Best precision, low recall (87.6 vs 82.3, 65.8 vs 74.8)
- Combine with method with best recall
- Priority to our system for non-null labels, other labels if we output null
- Results: 80.5% precision, 81.4% recall, 81.0% F-measure
- Better than two systems individually
- Combination of two best systems yields an average between the two (hence not as good as the better of the two)

Current Work

- Improve recall by synchronous assignments (Conll shared task 2008)
- Application to dialogue systems for English and French

Relevance for Linguistics

- Linking theory
- Model of implicit learning and knowledge of language
- Methodology