Dependency Grammars and Parsers

Deep Processing for NLP
Ling571
January 28, 2015

Roadmap

- PCFGs: Efficiencies and Reranking
- Dependency Grammars
  - Definition
  - Motivation:
    - Limitations of Context-Free Grammars
- Dependency Parsing
  - By conversion to CFG
  - By Graph-based models
  - By transition-based parsing

Efficiency

- PCKY is $|G|n^3$
  - Grammar can be huge
  - Grammar can be extremely ambiguous
    - 100s of analyses not unusual, esp. for long sentences
- However, only care about best parses
  - Others can be pretty bad
- Can we use this to improve efficiency?

Beam Thresholding

- Inspired by beam search algorithm
- Assume low probability partial parses unlikely to yield high probability overall
  - Keep only top k most probably partial parses
    - Retain only k choices per cell
      - For large grammars, could be 50 or 100
      - For small grammars, 5 or 10

Heuristic Filtering

- Intuition: Some rules/partial parses are unlikely to end up in best parse. Don't store those in table.

Heuristic Filtering

- Intuition: Some rules/partial parses are unlikely to end up in best parse. Don't store those in table.

- Exclusions:
  - Low frequency: exclude singleton productions

Heuristic Filtering

- Intuition: Some rules/partial parses are unlikely to end up in best parse. Don't store those in table.

- Exclusions:
  - Low frequency: exclude singleton productions
  - Low probability: exclude constituents x s.t. $p(x) < 10^{-200}$

Heuristic Filtering

- Intuition: Some rules/partial parses are unlikely to end up in best parse. Don't store those in table.

- Exclusions:
  - Low frequency: exclude singleton productions
  - Low probability: exclude constituents x s.t. $p(x) < 10^{-200}$
  - Low relative probability:
    - Exclude x if there exists y s.t. $p(y) > 100 * p(x)$

Reranking

- Issue: Locality
  - PCFG probabilities associated with rewrite rules
  - Context-free grammars

Reranking

- Issue: Locality
  - PCFG probabilities associated with rewrite rules
  - Context-free grammars
  - Approaches create new rules incorporating context:
    - Parent annotation, Markovization, lexicalization
  - Other problems:

Reranking

- Issue: Locality
  - PCFG probabilities associated with rewrite rules
  - Context-free grammars
  - Approaches create new rules incorporating context:
    - Parent annotation, Markovization, lexicalization
  - Other problems:
    - Increase rules, sparseness
- Need approach that incorporates broader, global info

Discriminative Parse Reranking

- General approach:
  - Parse using (L)PCFG
  - Obtain top-N parses
  - Re-rank top-N parses using better features

Discriminative Parse Reranking

- General approach:
  - Parse using (L)PCFG
  - Obtain top-N parses
  - Re-rank top-N parses using better features

- Discriminative reranking
  - Use arbitrary features in reranker (MaxEnt)
    - E.g. right-branching-ness, speaker identity, conjunctive parallelism, fragment frequency, etc

Reranking Effectiveness

- How can reranking improve?
  - N-best includes the correct parse
- Estimate maximum improvement
  - **Oracle** parse selection
    - Selects correct parse from N-best
      - If it appears
- E.g. Collins parser (2000)
  - Base accuracy: 0.897
  - Oracle accuracy on 50-best: 0.968
- Discriminative reranking: 0.917

Dependency Grammar

- CFGs:
  - Phrase-structure grammars
  - Focus on modeling constituent structure

Dependency Grammar

- CFGs:
  - Phrase-structure grammars
  - Focus on modeling constituent structure

- Dependency grammars:
  - Syntactic structure described in terms of
    - Words
    - Syntactic/Semantic relations between words

Dependency Parse

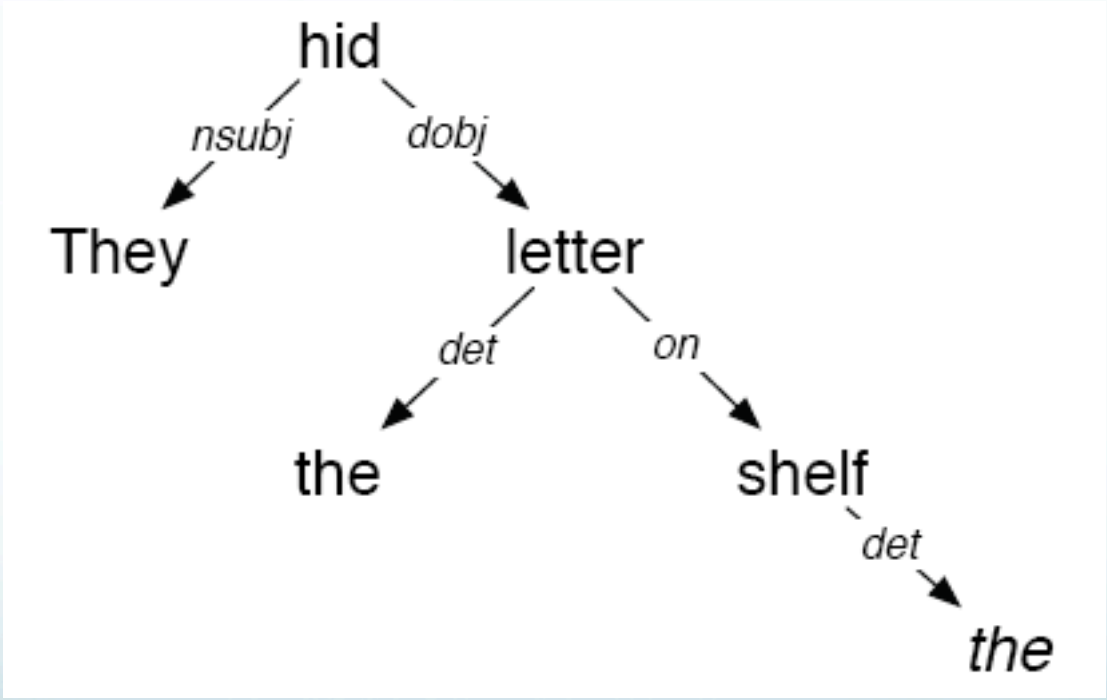
- A dependency parse is a tree, where
  - Nodes correspond to words in utterance
  - Edges between nodes represent dependency relations
    - Relations may be labeled (or not)

Dependency Relations

Argument Dependencies	Description
nsubj	nominal subject
csubj	clausal subject
dobj	direct object
iobj	indirect object
pobj	object of preposition
Modifier Dependencies	**Description**
tmod	temporal modifier
appos	appositional modifier
det	determiner
prep	prepositional modifier

Dependency Parse Example

- They hid the letter on the shelf



Why Dependency Grammar?

- More natural representation for many tasks
  - Clear encapsulation of predicate-argument structure
    - Phrase structure may obscure, e.g. wh-movement

Why Dependency Grammar?

- More natural representation for many tasks
  - Clear encapsulation of predicate-argument structure
    - Phrase structure may obscure, e.g. wh-movement
  - Good match for question-answering, relation extraction
    - Who did what to whom
    - Build on parallelism of relations between question/relation specifications and answer sentences

Why Dependency Grammar?

- Easier handling of flexible or free word order
  - How does CFG handle variations in word order?

Why Dependency Grammar?

- Easier handling of flexible or free word order
  - How does CFG handle variations in word order?
    - Adds extra phrases structure rules for alternatives
      - Minor issue in English, explosive in other langs
  - What about dependency grammar?

Why Dependency Grammar?

- Easier handling of flexible or free word order
  - How does CFG handle variations in word order?
    - Adds extra phrases structure rules for alternatives
      - Minor issue in English, explosive in other langs
  - What about dependency grammar?
    - No difference: link represents relation
    - Abstracts away from surface word order

Why Dependency Grammar?

- Natural efficiencies:
  - CFG: Must derive full trees of many non-terminals

Why Dependency Grammar?

- Natural efficiencies:
  - CFG: Must derive full trees of many non-terminals
  - Dependency parsing:
    - For each word, must identify
      - Syntactic head, h
      - Dependency label, d

Why Dependency Grammar?

- Natural efficiencies:
  - CFG: Must derive full trees of many non-terminals
  - Dependency parsing:
    - For each word, must identify
      - Syntactic head, h
      - Dependency label, d
    - Inherently lexicalized
      - Strong constraints hold between pairs of words

Summary

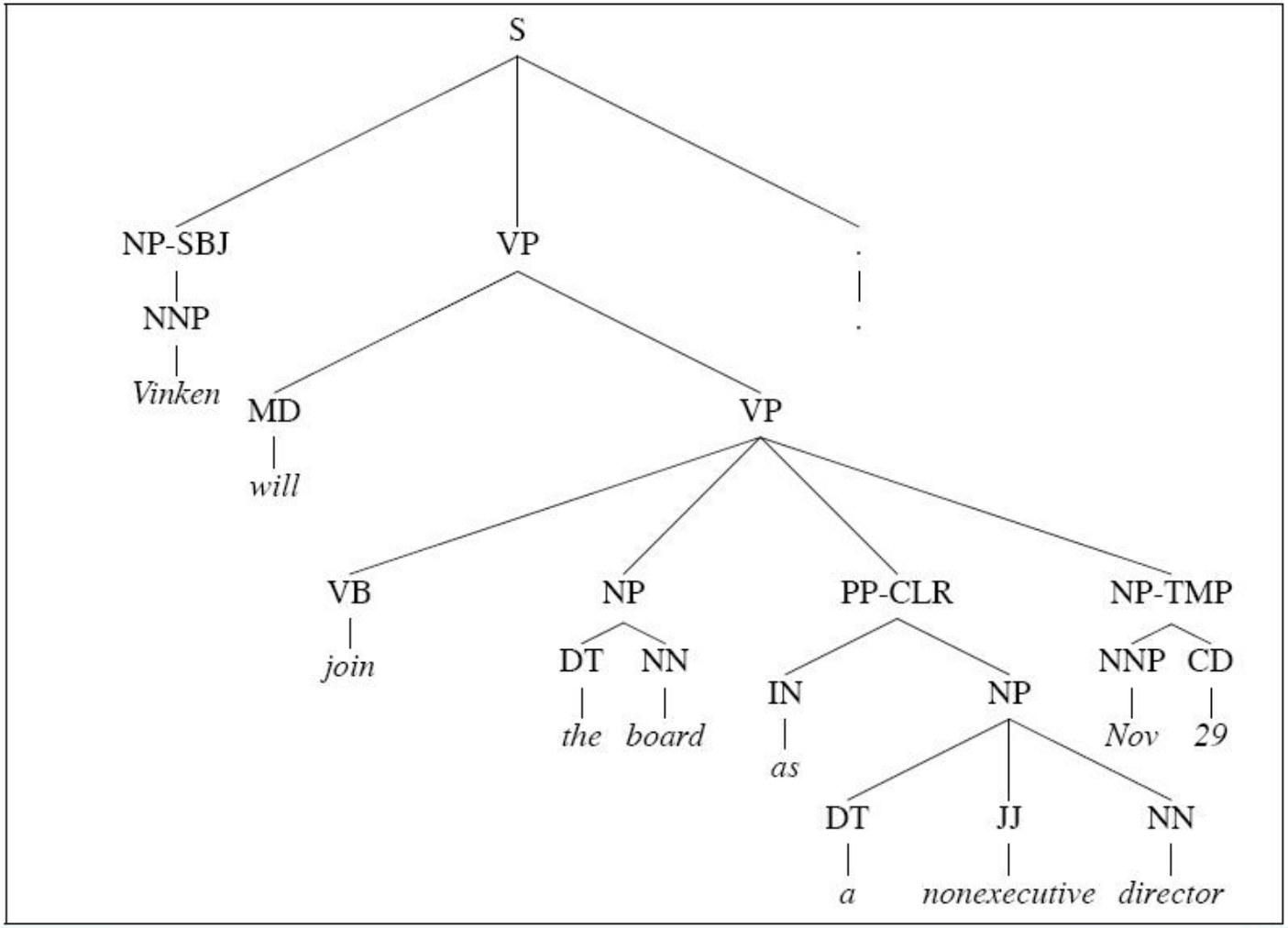
- Dependency grammar balances complexity and expressiveness
  - Sufficiently expressive to capture predicate-argument structure
  - Sufficiently constrained to allow efficient parsing

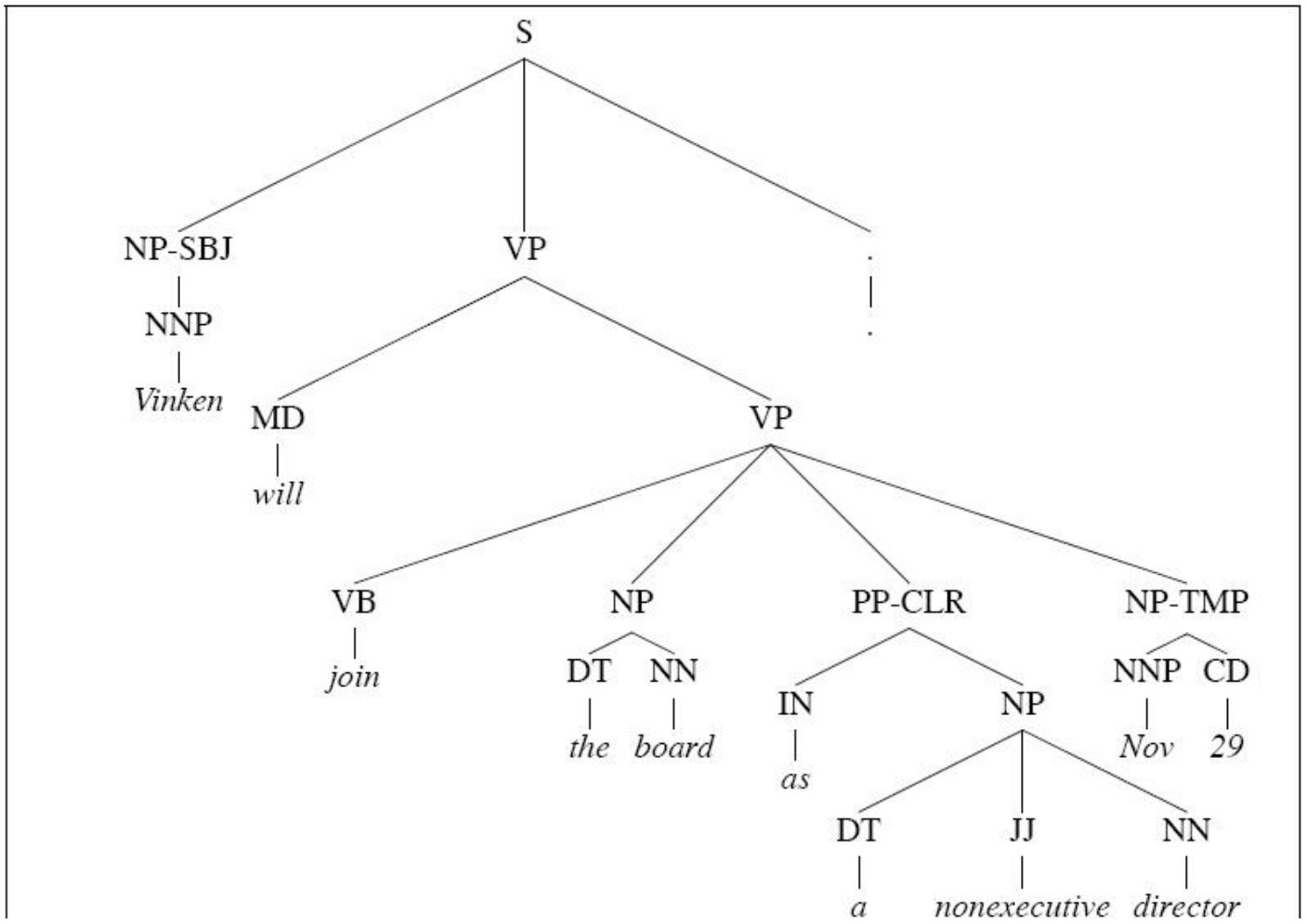
Conversion

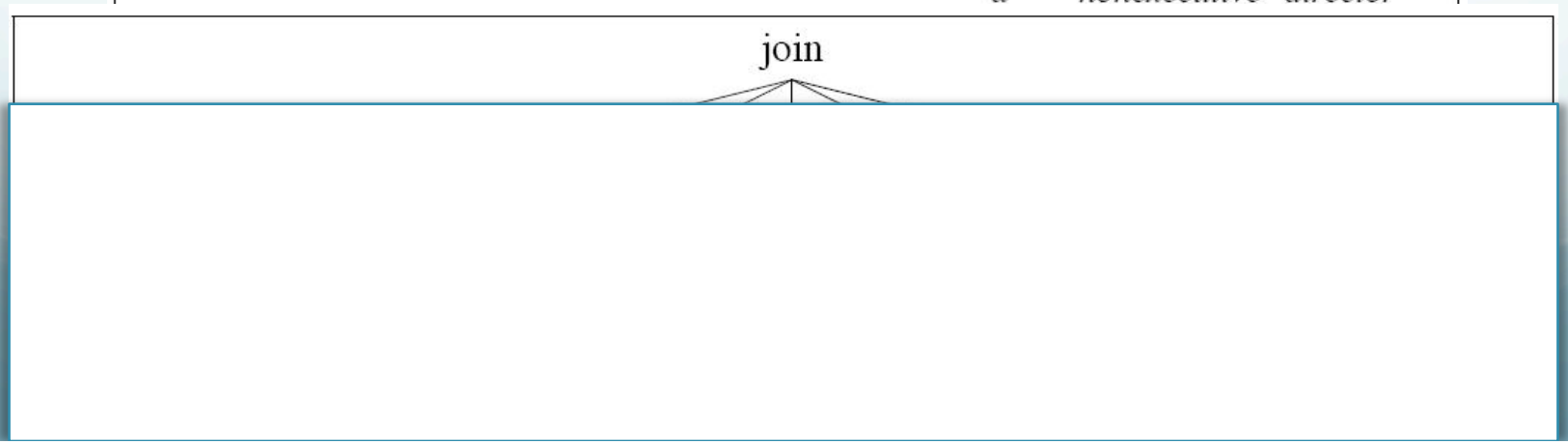
- Can convert phrase structure to dependency trees
  - Unlabeled dependencies

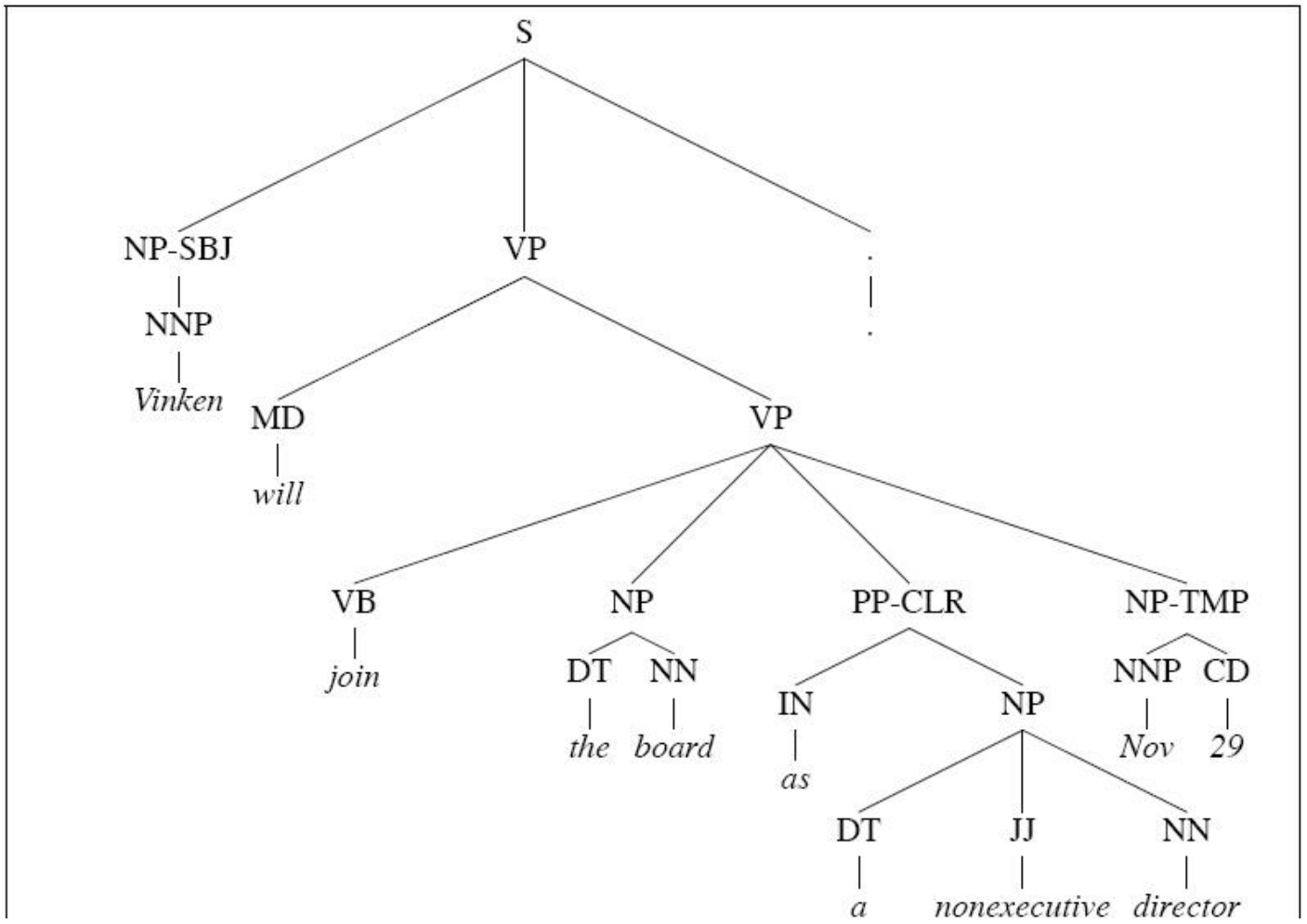
Conversion

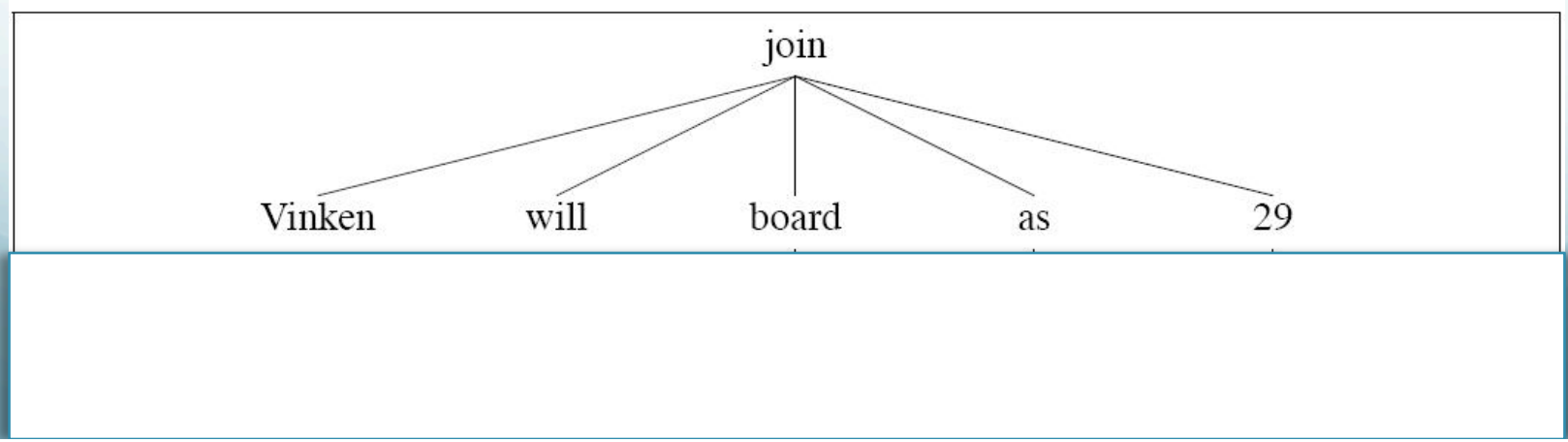
- Can convert phrase structure to dependency trees
  - Unlabeled dependencies
- Algorithm:
  - Identify all head children in PS tree
  - Make head of each non-head-child depend on head of head-child

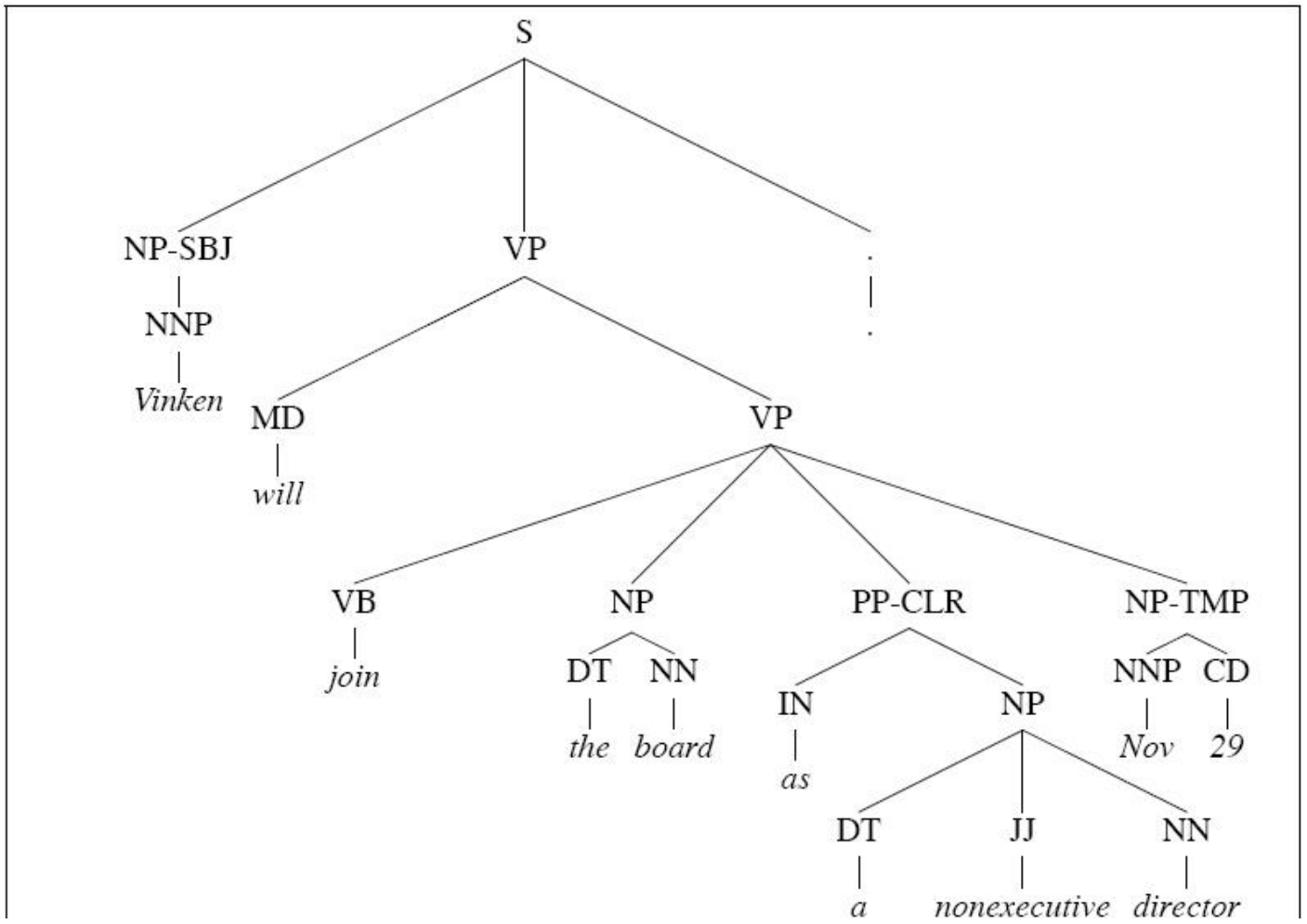

S
NP-SBJ
VP
.
NNP
.
Vinken
MD
VP
will
VB
NP
PP-CLR
NP-TMP
join
DT
NN
IN
NP
NNP
CD
the
board
as
Nov
29
DT
JJ
NN
a
nonexecutive
director

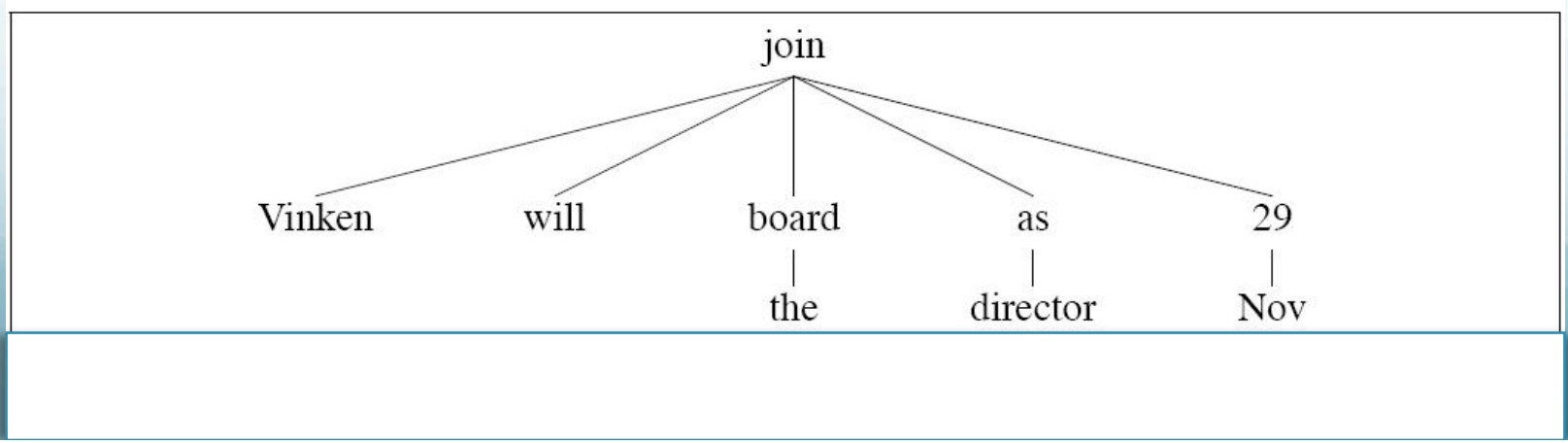

S
NP-SBJ
VP
.
NNP
.
Vinken
MD
VP
will
VB
NP
PP-CLR
NP-TMP
join
DT
NN
NNP
CD
the
board
IN
NP
Nov
29
as
DT
JJ
NN
a
nonexecutive
director

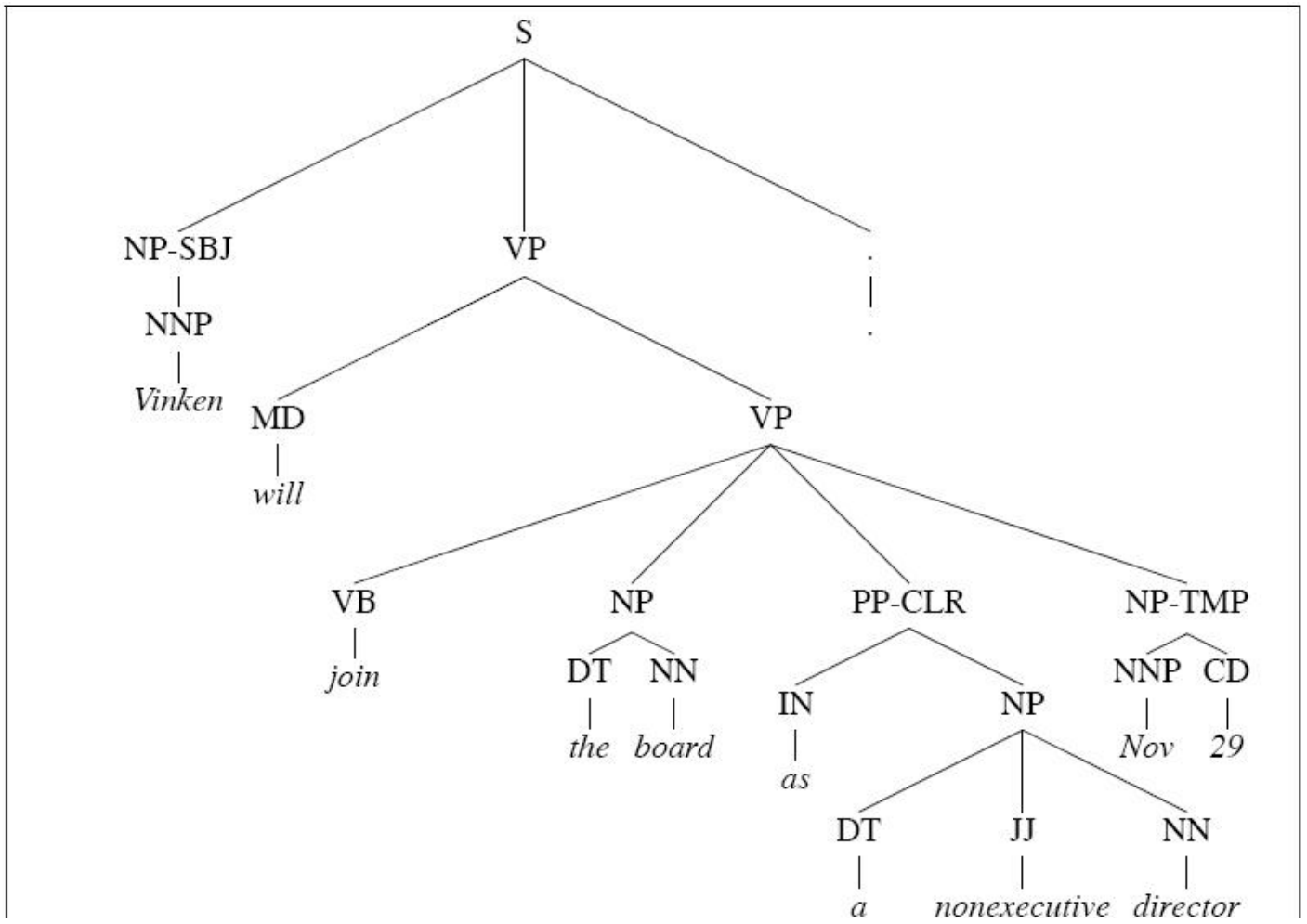

join

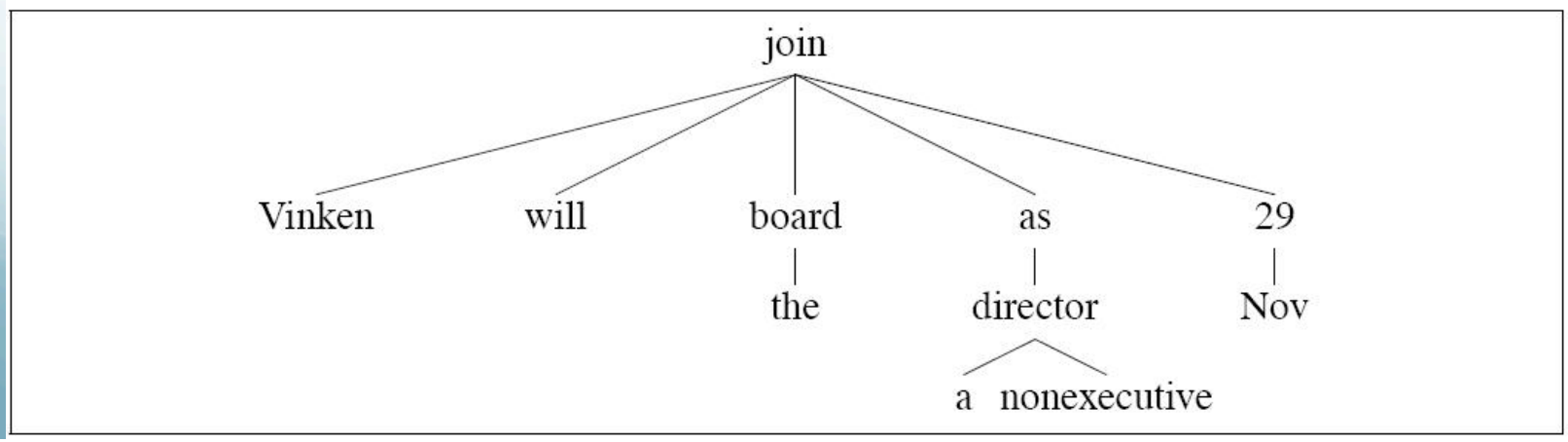

S
NP-SBJ
VP
.
NNP
Vinken
MD
VP
.
will
VB
NP
PP-CLR
NP-TMP
join
DT
NN
IN
NP
NNP
CD
the
board
as
Nov
29
DT
JJ
NN
a
nonexecutive
director


join
Vinken
will
board
as
29


S
NP-SBJ
VP
.
NNP
Vinken
MD
VP
will
VB
NP
PP-CLR
NP-TMP
join
DT
NN
IN
NP
NNP
CD
the
board
as
Nov
29
DT
JJ
NN
a
nonexecutive
director


join
Vinken
will
board
as
29
the
director
Nov


S
NP-SBJ
VP
.
NNP
Vinken
MD
VP
will
VB
NP
PP-CLR
NP-TMP
join
DT
NN
IN
NP
NNP
CD
the
board
as
Nov
29
DT
JJ
NN
a
nonexecutive
director

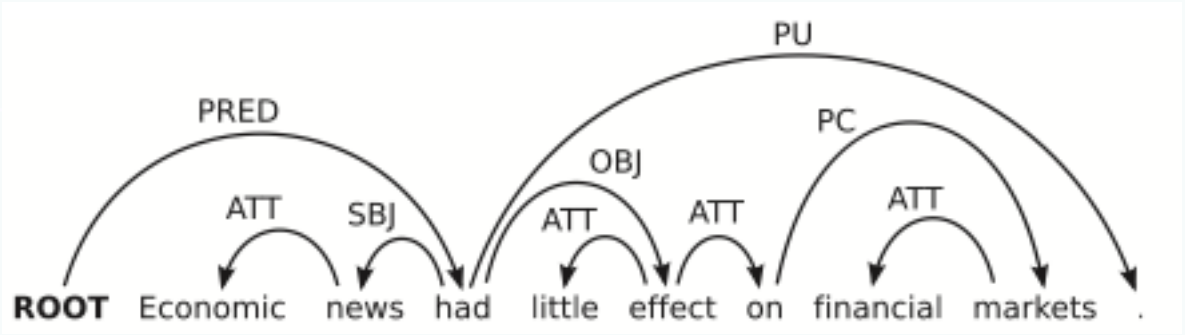

join
Vinken
will
board
as
29
the
director
Nov
a
nonexecutive

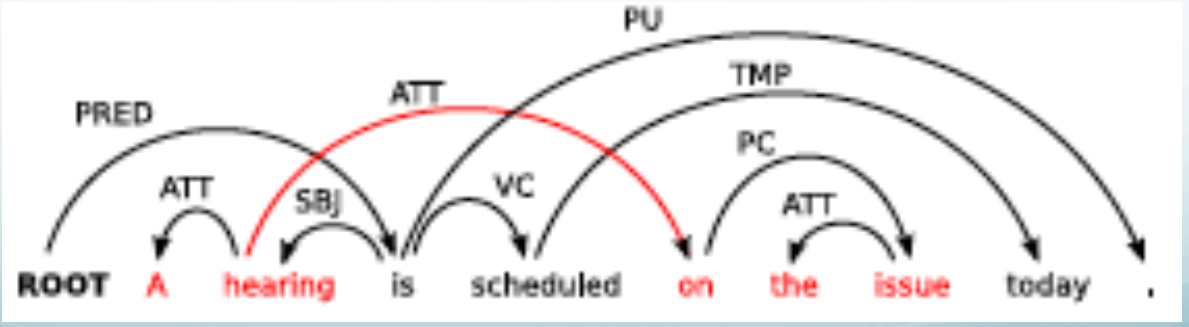
Dependency Parsing

- Three main strategies:
  - Convert dependency trees to PS trees
    - Parse using standard algorithms $O(n^3)$
  - Employ graph-based optimization
    - Weights learned by machine learning
  - Shift-reduce approaches based on current word/state
    - Attachment based on machine learning

Parsing by PS Conversion

- Can map any projective dependency tree to PS tree
  - Non-terminals indexed by words
    - "Projective": no crossing dependency arcs for ordered words



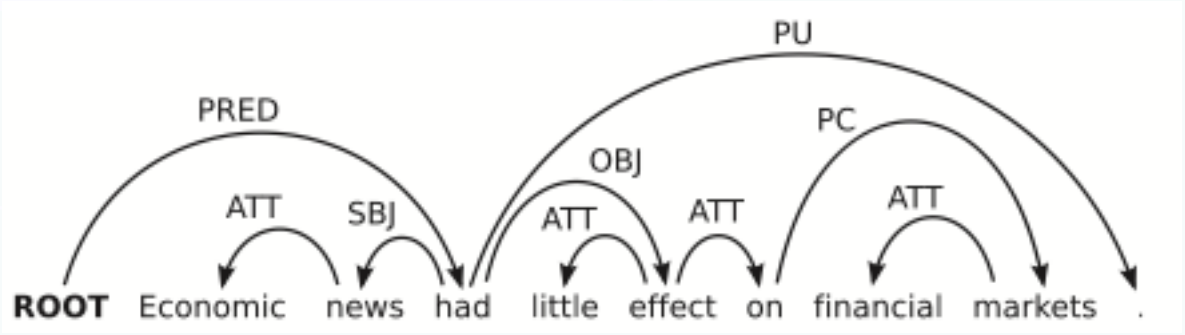


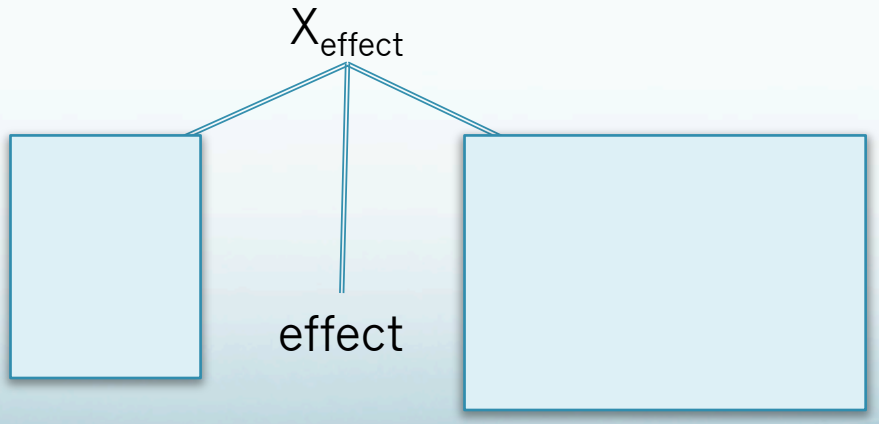
Dep to PS Tree Conversion

- For each node *w* with outgoing arcs,
  - Convert the subtree *w* and its dependents $t_1,..,t_n$ to
  - New subtree rooted at X_w with child *w* and
    - Subtrees at $t_1,..,t_n$ in the original sentence order

Dep to PS Tree Conversion

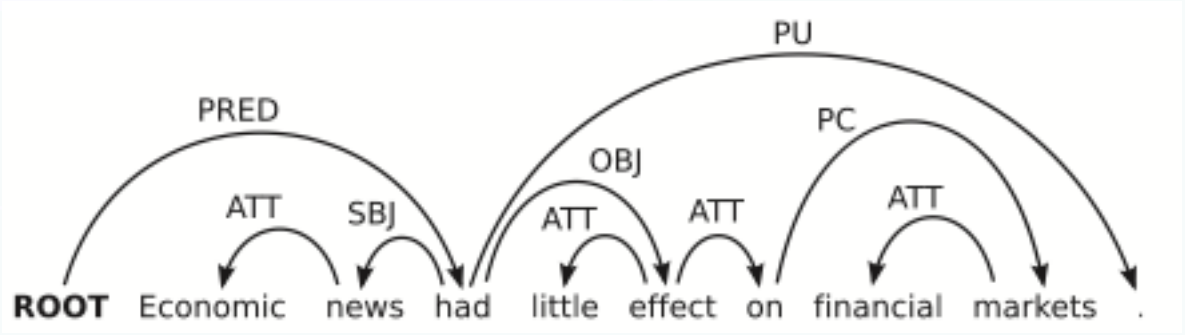
E.g., for ‘effect’

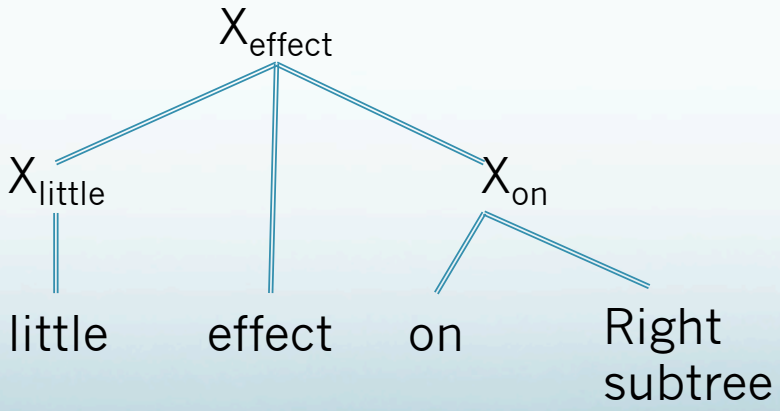




Dep to PS Tree Conversion

E.g., for ‘effect’





PS to Dep Tree Conversion

- What about the dependency labels?
  - Attach labels to non-terminals associated with non-heads
  - E.g. X_{little} ➔ $X_{little:nmod}$

PS to Dep Tree Conversion

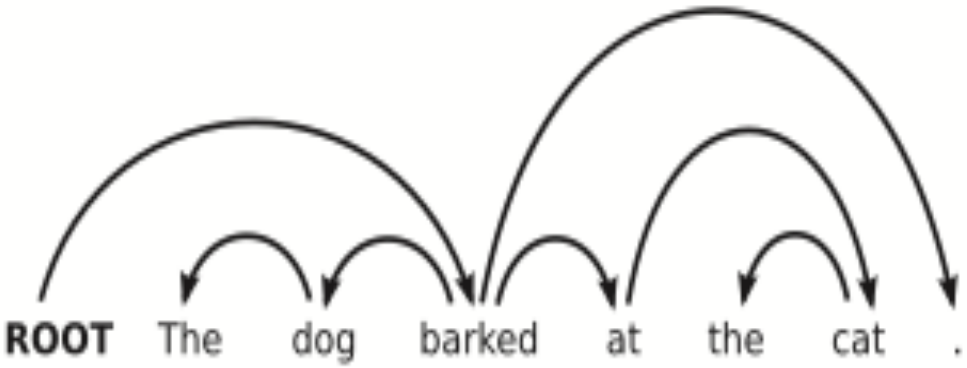
- What about the dependency labels?
  - Attach labels to non-terminals associated with non-heads
  - E.g. X_{little} ➔ $X_{little:nmod}$

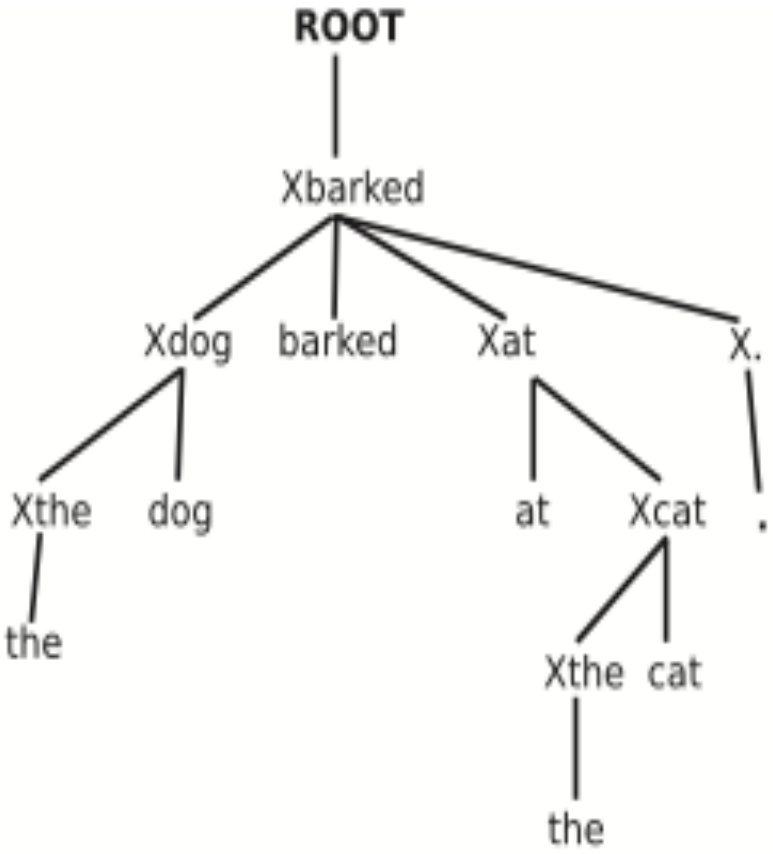
- Doesn't create typical PS trees
  - Does create fully lexicalized, context-free trees
    - Also labeled

PS to Dep Tree Conversion

- What about the dependency labels?
  - Attach labels to non-terminals associated with non-heads
  - E.g. X_{little} ➔ $X_{little:nmod}$
- Doesn't create typical PS trees
  - Does create fully lexicalized, context-free trees
    - Also labeled
- Can be parsed with any standard CFG parser
  - E.g. CKY, Earley

Full Example Trees





Example from J. Moore, 2013

Graph-based Dependency Parsing

- Goal: Find the highest scoring dependency tree T for sentence S
  - If S is unambiguous, T is the correct parse.
  - If S is ambiguous, T is the highest scoring parse.

Graph-based Dependency Parsing

- Goal: Find the highest scoring dependency tree T for sentence S
  - If S is unambiguous, T is the correct parse.
  - If S is ambiguous, T is the highest scoring parse.
- Where do scores come from?
  - Weights on dependency edges by machine learning
  - Learned from large dependency treebank

Graph-based Dependency Parsing

- Goal: Find the highest scoring dependency tree T for sentence S
  - If S is unambiguous, T is the correct parse.
  - If S is ambiguous, T is the highest scoring parse.
- Where do scores come from?
  - Weights on dependency edges by machine learning
  - Learned from large dependency treebank
- Where are the grammar rules?

Graph-based Dependency Parsing

- Goal: Find the highest scoring dependency tree T for sentence S
  - If S is unambiguous, T is the correct parse.
  - If S is ambiguous, T is the highest scoring parse.
- Where do scores come from?
  - Weights on dependency edges by machine learning
  - Learned from large dependency treebank
- Where are the grammar rules?
  - There aren't any; data-driven processing

Graph-based Dependency Parsing

- Map dependency parsing to maximum spanning tree

Graph-based Dependency Parsing

- Map dependency parsing to maximum spanning tree
- Idea:
  - Build initial graph: fully connected
    - Nodes: words in sentence to parse

Graph-based Dependency Parsing

- Map dependency parsing to maximum spanning tree
- Idea:
  - Build initial graph: fully connected
    - Nodes: words in sentence to parse
    - Edges: Directed edges between all words
      - + Edges from ROOT to all words

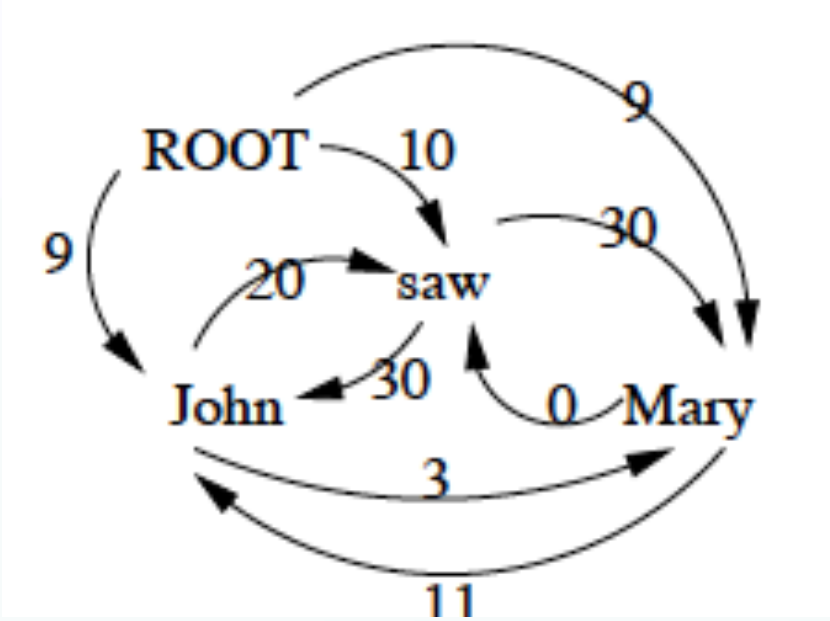
Graph-based Dependency Parsing

- Map dependency parsing to maximum spanning tree
- Idea:
  - Build initial graph: fully connected
    - Nodes: words in sentence to parse
    - Edges: Directed edges between all words
      - + Edges from ROOT to all words
  - Identify maximum spanning tree
    - Tree s.t. all nodes are connected
    - Select such tree with highest weight

Graph-based Dependency Parsing

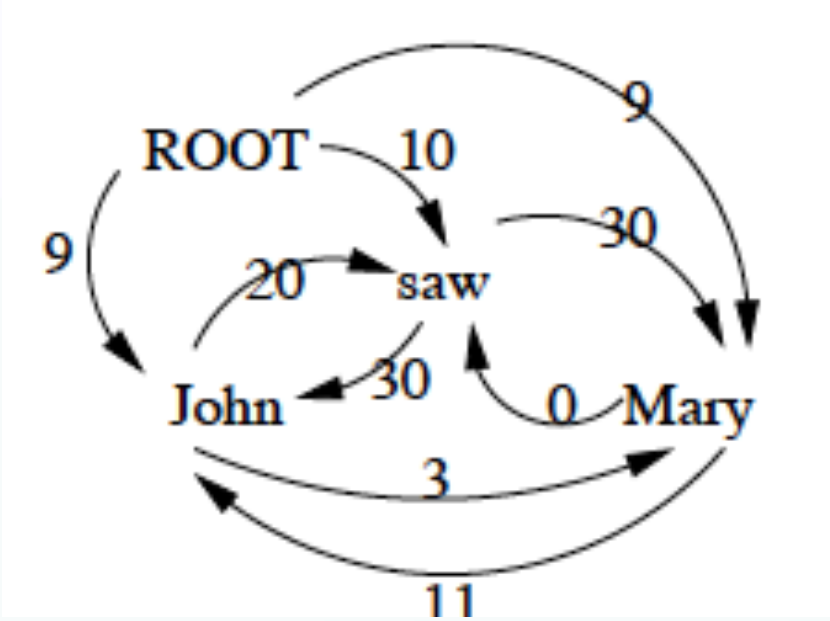
- Map dependency parsing to maximum spanning tree
- Idea:
  - Build initial graph: fully connected
    - Nodes: words in sentence to parse
    - Edges: Directed edges between all words
      - + Edges from ROOT to all words
  - Identify maximum spanning tree
    - Tree s.t. all nodes are connected
    - Select such tree with highest weight
    - Arc-factored model: Weights depend on end nodes & link
      - Weight of tree is sum of participating arcs

Initial Tree



- Sentence: John saw Mary (McDonald et al, 2005)
  - All words connected; ROOT only has outgoing arcs

Initial Tree



- Sentence: John saw Mary (McDonald et al, 2005)
  - All words connected; ROOT only has outgoing arcs
- Goal: Remove arcs to create a tree covering all words
  - Resulting tree is dependency parse

Maximum Spanning Tree

- McDonald et al, 2005 use variant of Chu-Liu-Edmonds algorithm for MST (CLE)

Maximum Spanning Tree

- McDonald et al, 2005 use variant of Chu-Liu-Edmonds algorithm for MST (CLE)
- Sketch of algorithm:
  - For each node, greedily select incoming arc with max w
  - If the resulting set of arcs forms a tree, this is the MST.
  - If not, there must be a cycle.

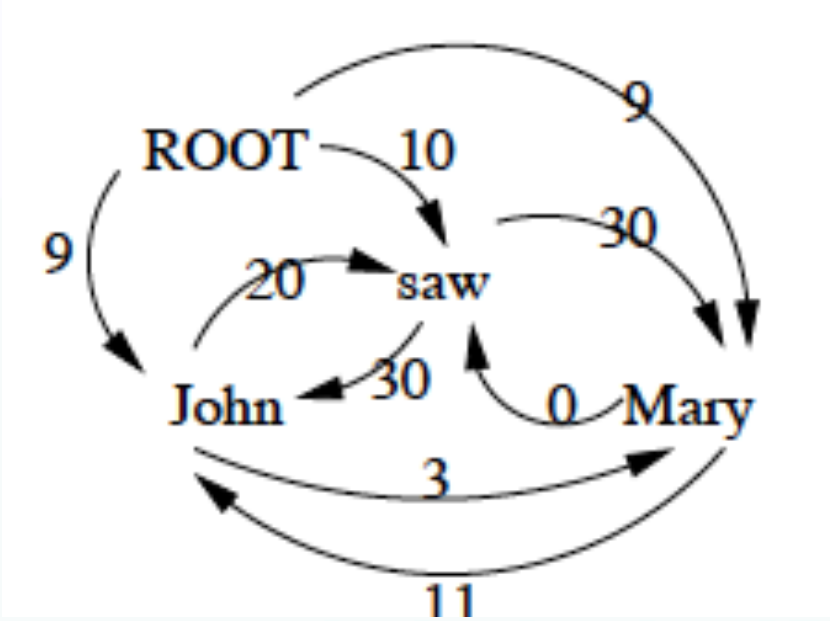
Maximum Spanning Tree

- McDonald et al, 2005 use variant of Chu-Liu-Edmonds algorithm for MST (CLE)
- Sketch of algorithm:
  - For each node, greedily select incoming arc with max w
  - If the resulting set of arcs forms a tree, this is the MST.
  - If not, there must be a cycle.
    - "Contract" the cycle: Treat it as a single vertex
    - Recalculate weights into/out of the new vertex
    - Recursively do MST algorithm on resulting graph

Maximum Spanning Tree

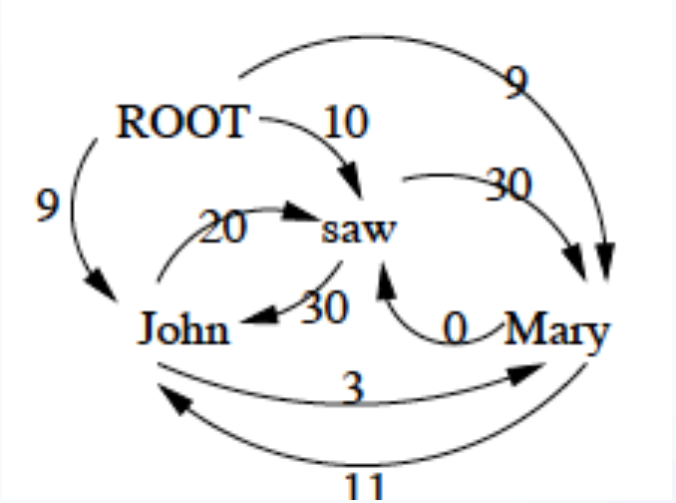
- McDonald et al, 2005 use variant of Chu-Liu-Edmonds algorithm for MST (CLE)
- Sketch of algorithm:
  - For each node, greedily select incoming arc with max w
  - If the resulting set of arcs forms a tree, this is the MST.
  - If not, there must be a cycle.
    - "Contract" the cycle: Treat it as a single vertex
    - Recalculate weights into/out of the new vertex
    - Recursively do MST algorithm on resulting graph
- Running time: naïve: $O(n^3)$; Tarjan: $O(n^2)$
  - Applicable to non-projective graphs

Initial Tree



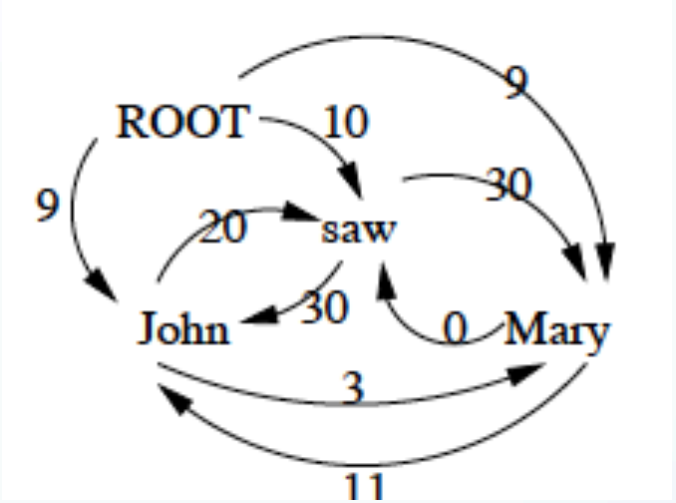
CLE: Step 1

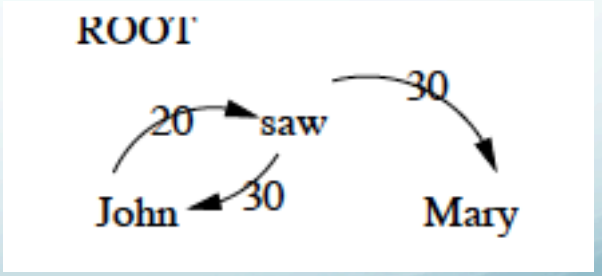
- Find maximum incoming arcs



CLE: Step 1

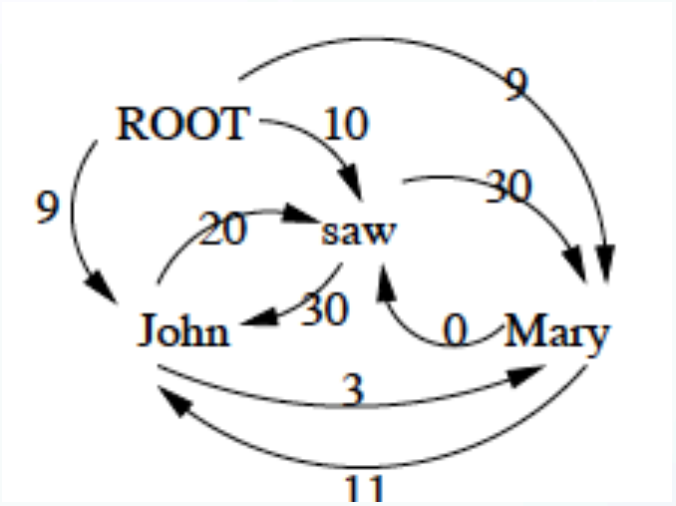
- Find maximum incoming arcs
  - Is the result a tree?

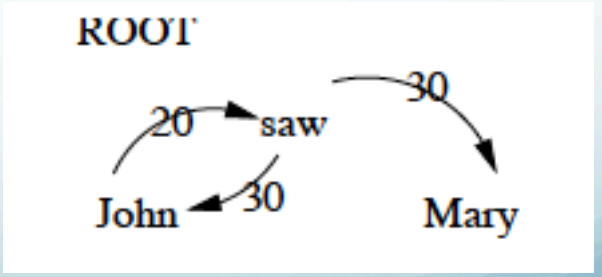




CLE: Step 1

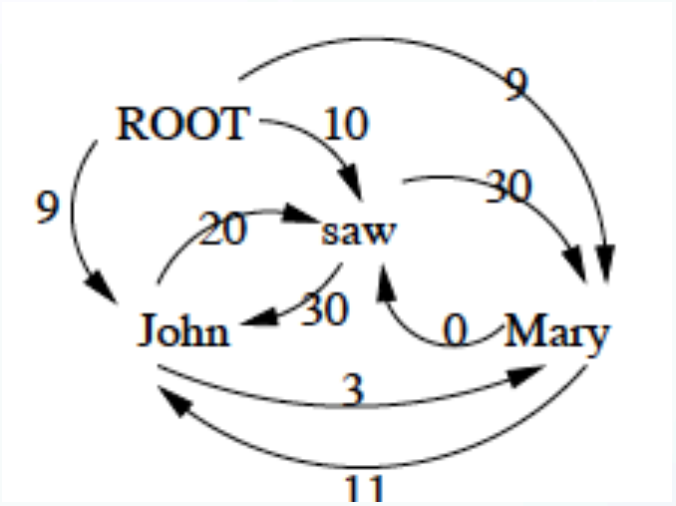
- Find maximum incoming arcs
  - Is the result a tree?
    - No
  - Is there a cycle?

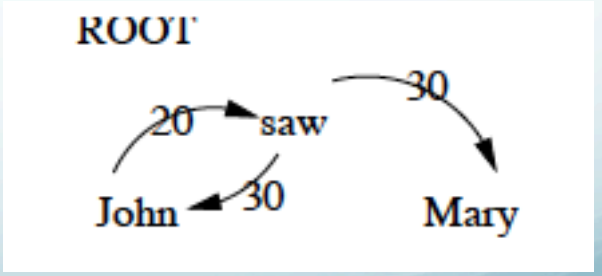




CLE: Step 1

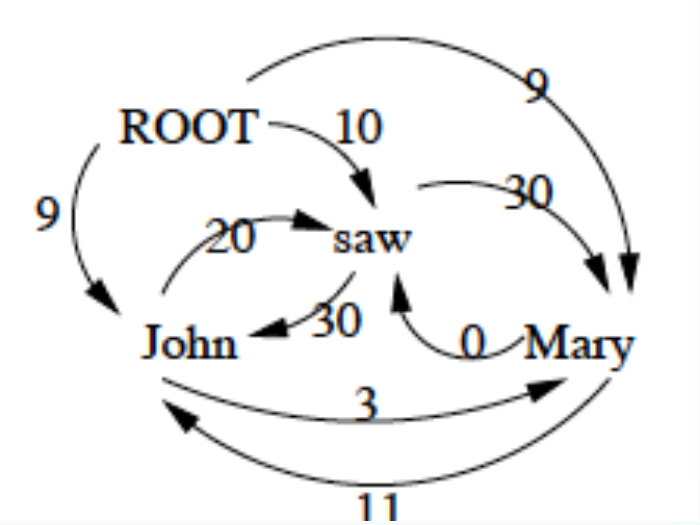
- Find maximum incoming arcs
  - Is the result a tree?
    - No
  - Is there a cycle?
    - Yes, John/saw





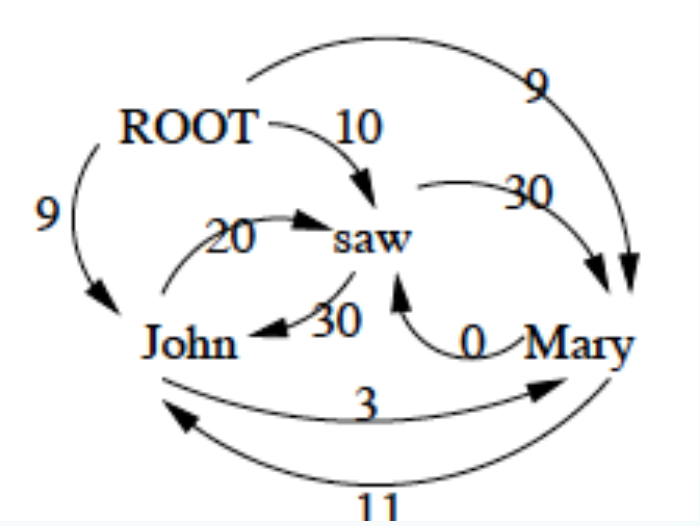
CLE: Step 2

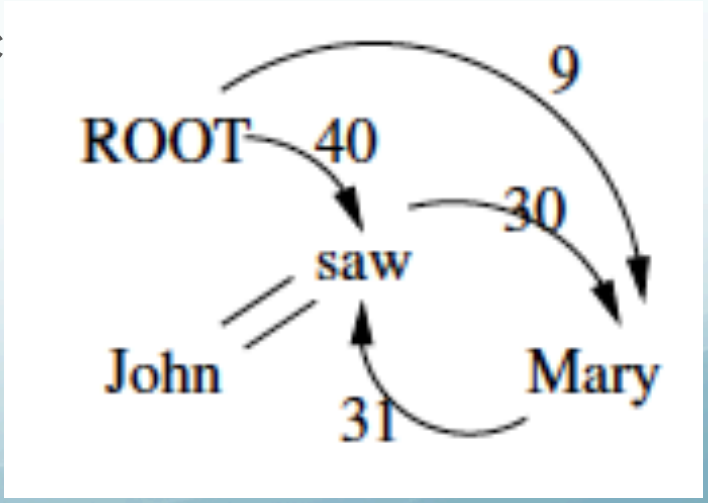
- Since there's a cycle:
  - Contract cycle & reweight
  - John+saw as single vertex



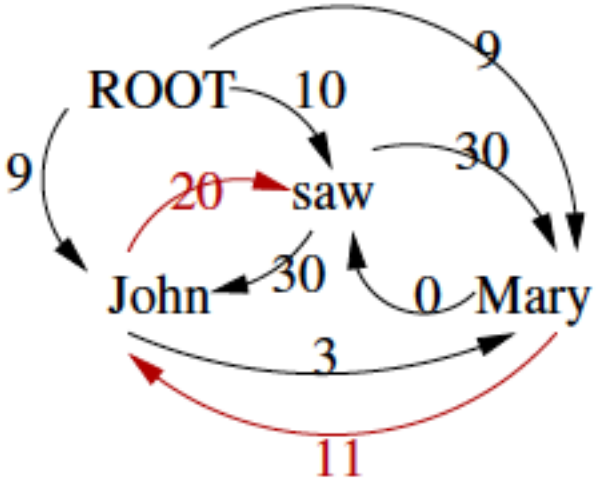
CLE: Step 2

- Since there's a cycle:
  - Contract cycle & reweight
  - John+saw as single vertex
  - Calculate weights in & out as:
    - Maximum based on internal arc
    - and original nodes
- Recurse

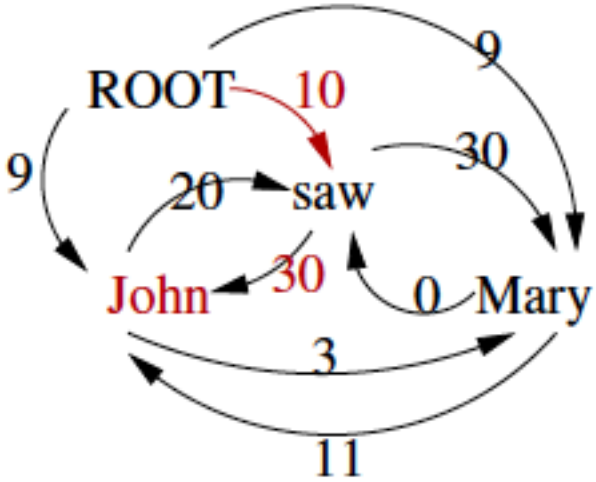




Calculating Graph



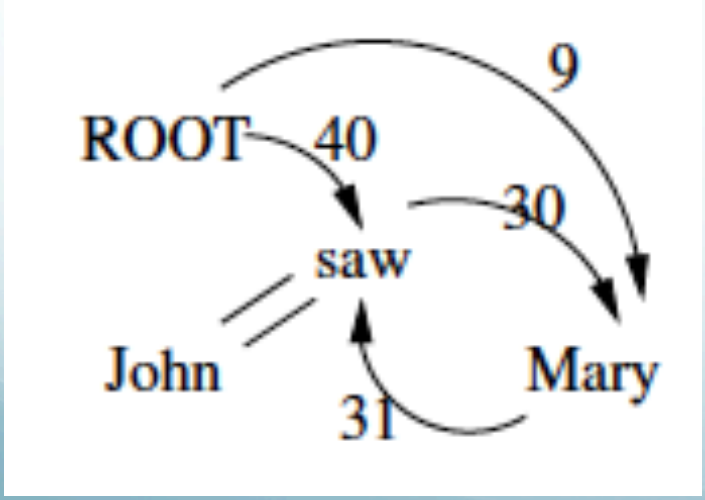
s(Mary, C) 11+20 = 31



s(ROOT, C) 10+30 = 40

CLE: Recursive Step

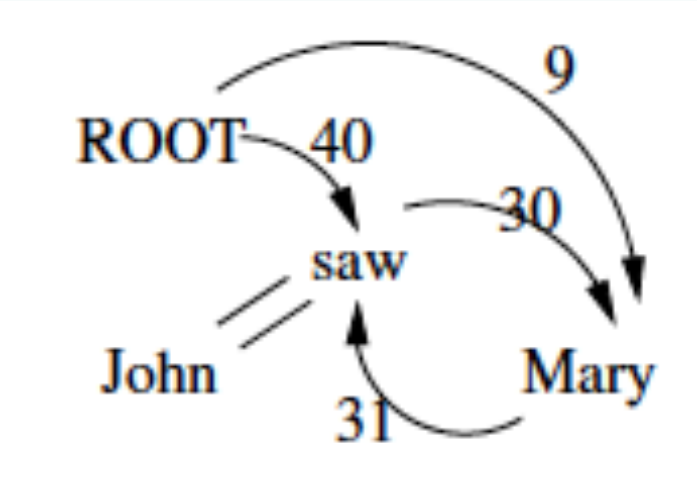
- In new graph, find graph of
  - Max weight incoming arc for each word

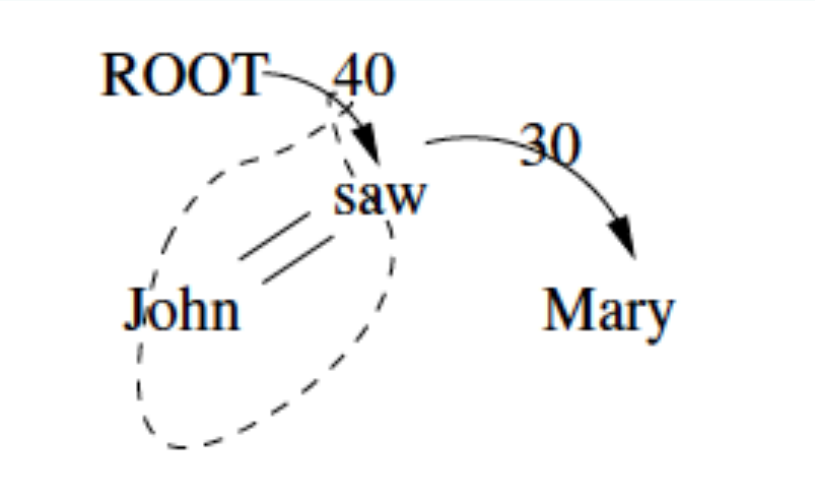


CLE: Recursive Step

- In new graph, find graph of
  - Max weight incoming arc for each word

- Is it a tree?

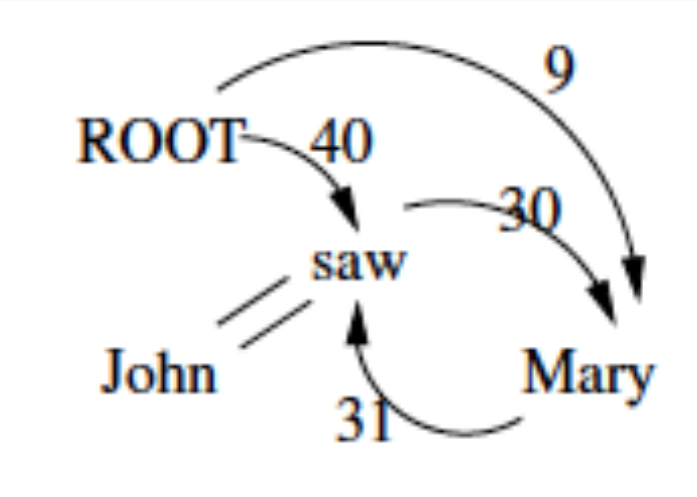


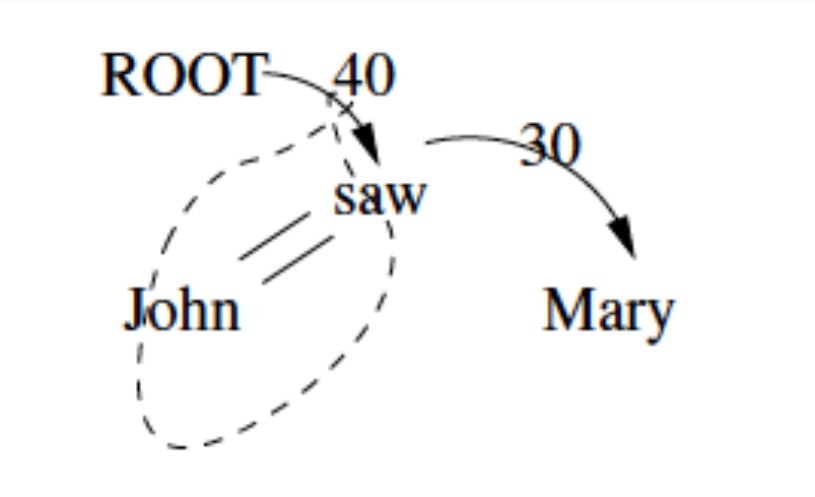


CLE: Recursive Step

- In new graph, find graph of
  - Max weight incoming arc for each word

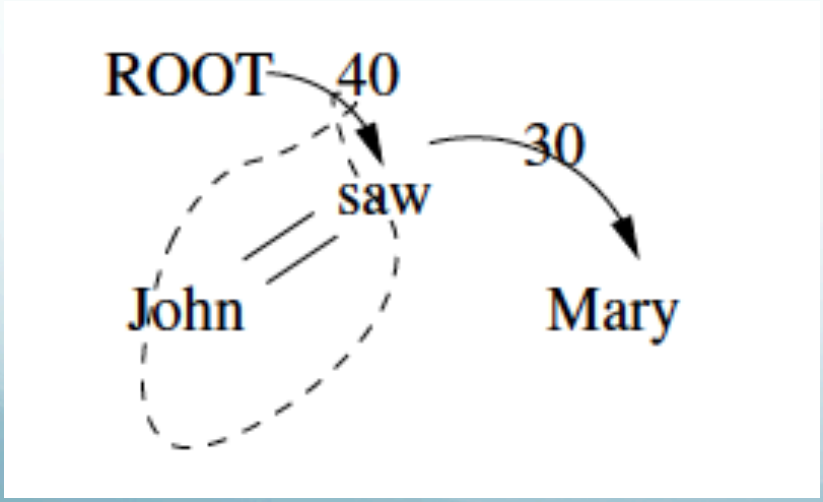
- Is it a tree? Yes!
  - MST, but must recover internal arcs ➔ parse





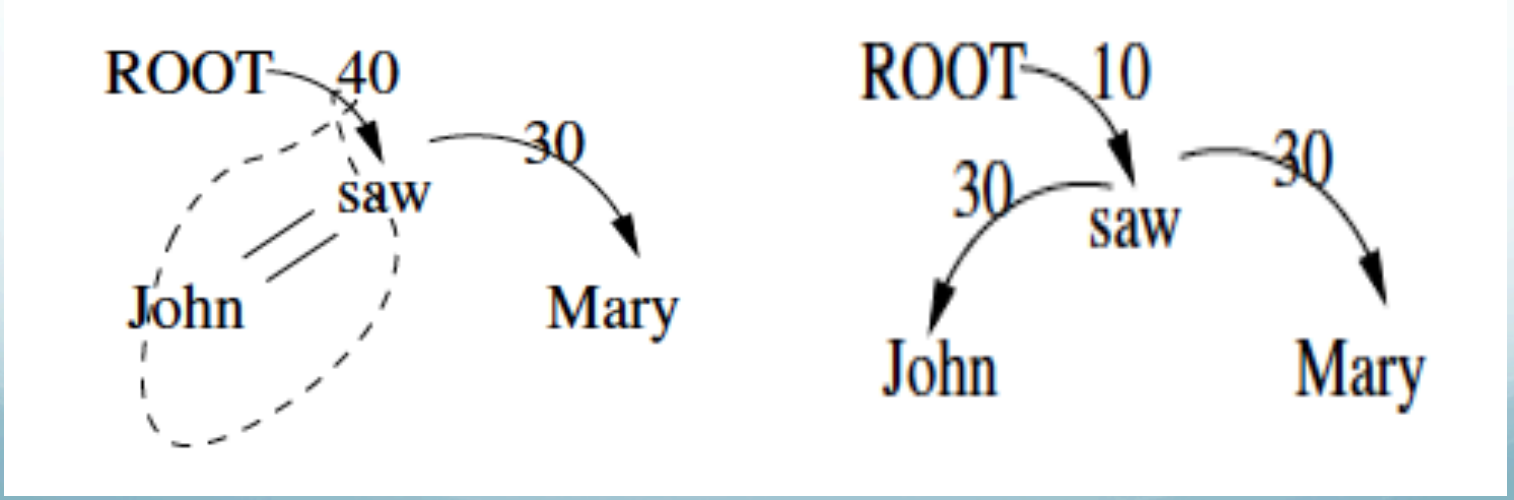
CLE: Recovering Graph

- Found maximum spanning tree
  - Need to ‘pop’ collapsed nodes
- Expand “ROOT → John+saw” = 40



CLE: Recovering Graph

- Found maximum spanning tree
  - Need to ‘pop’ collapsed nodes
- Expand “ROOT → John+saw” = 40
- MST and complete dependency parse



Learning Weights

- Weights for arc-factored model learned from corpus
  - Weights learned for tuple (w_i, w_j, l)

Learning Weights

- Weights for arc-factored model learned from corpus
  - Weights learned for tuple (w_i,w_j,l)
- McDonald et al, 2005 employed discriminative ML
  - Perceptron algorithm or large margin variant

Learning Weights

- Weights for arc-factored model learned from corpus
  - Weights learned for tuple (w_i,w_j,l)
- McDonald et al, 2005 employed discriminative ML
  - Perceptron algorithm or large margin variant
- Operates on vector of local features

Features for Learning Weights

- Simple categorical features for (w_i, L, w_j) including:
  - Identity of w_i (or char 5-gram prefix), POS of w_i
  - Identity of w_j (or char 5-gram prefix), POS of w_j
  - Label of L, direction of L
  - Sequence of POS tags b/t w_i, w_j
  - Number of words b/t w_i, w_j
  - POS tag of w_{i-1}, POS tag of w_{i+1}
  - POS tag of w_{j-1}, POS tag of w_{j+1}
- Features conjoined with direction of attachment and distance b/t words

Dependency Parsing

- Dependency grammars:
  - Compactly represent pred-arg structure
  - Lexicalized, localized
  - Natural handling of flexible word order
- Dependency parsing:
  - Conversion to phrase structure trees
  - Graph-based parsing (MST), efficient non-proj $O(n^2)$
  - Transition-based parser
    - MALTparser: very efficient O(n)
      - Optimizes local decisions based on many rich features