Research in Natural Language Processing

Daniel Gildea

Problems in Natural Language

- Part of Speech Tagging
- Parsing
- Information Extraction
- Generation
- Question Answering
- Summarization
- Dialog Systems
- Machine Translation

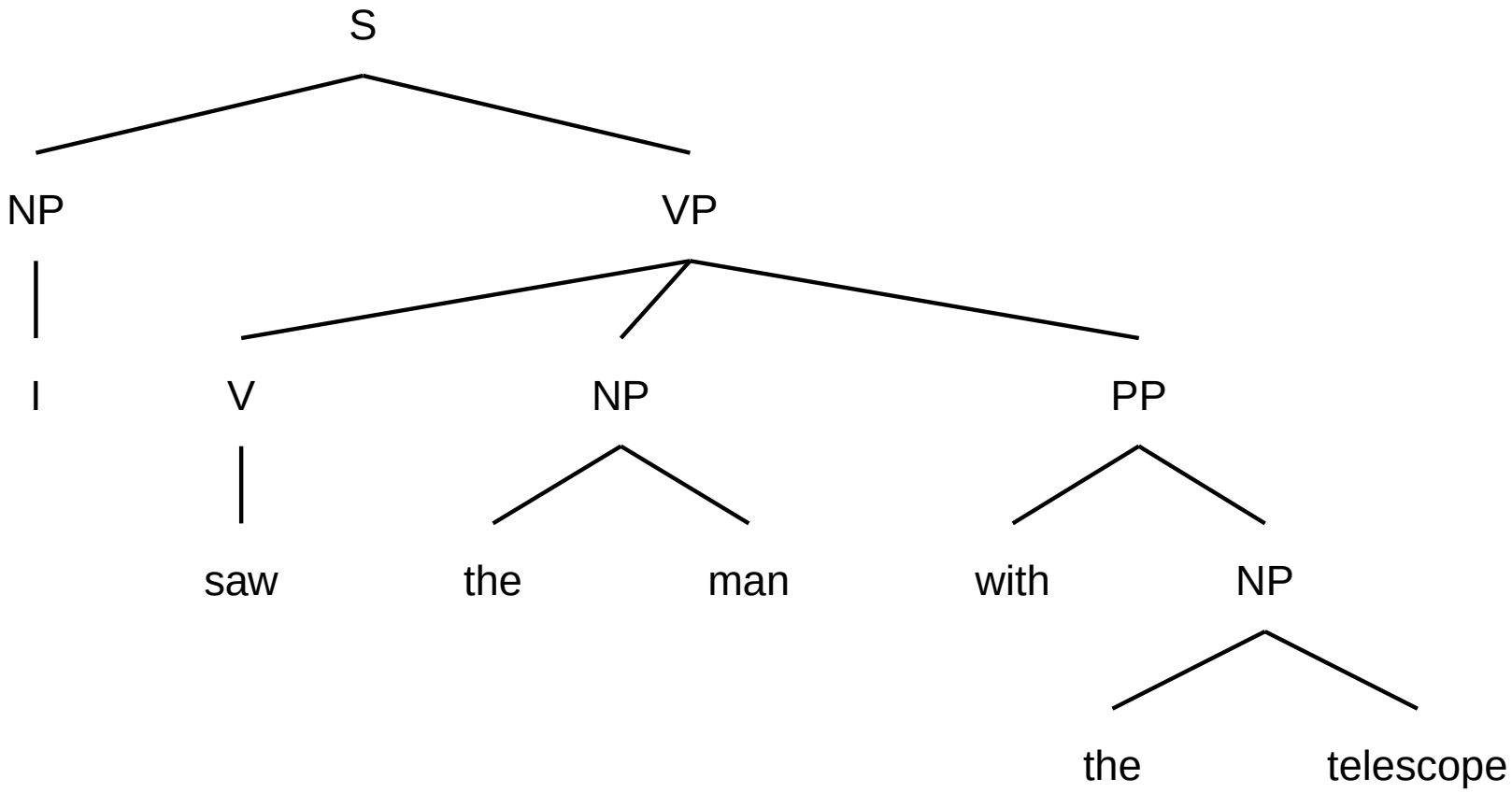
Basic Problem: Resolving Ambiguity

- Time flies like an arrow.
- Time flies with a stopwatch.
- I saw the man with the telescope.

Approaches to Resolving Ambiguity

- World Knowledge: Telescope is used to make things look bigger.
- Statistical Techniques: learn from examples
  - but: need examples that machines can understand
  - or: need machines that can guess from examples!
- This talk: language understanding, and translation

Structural Ambiguity

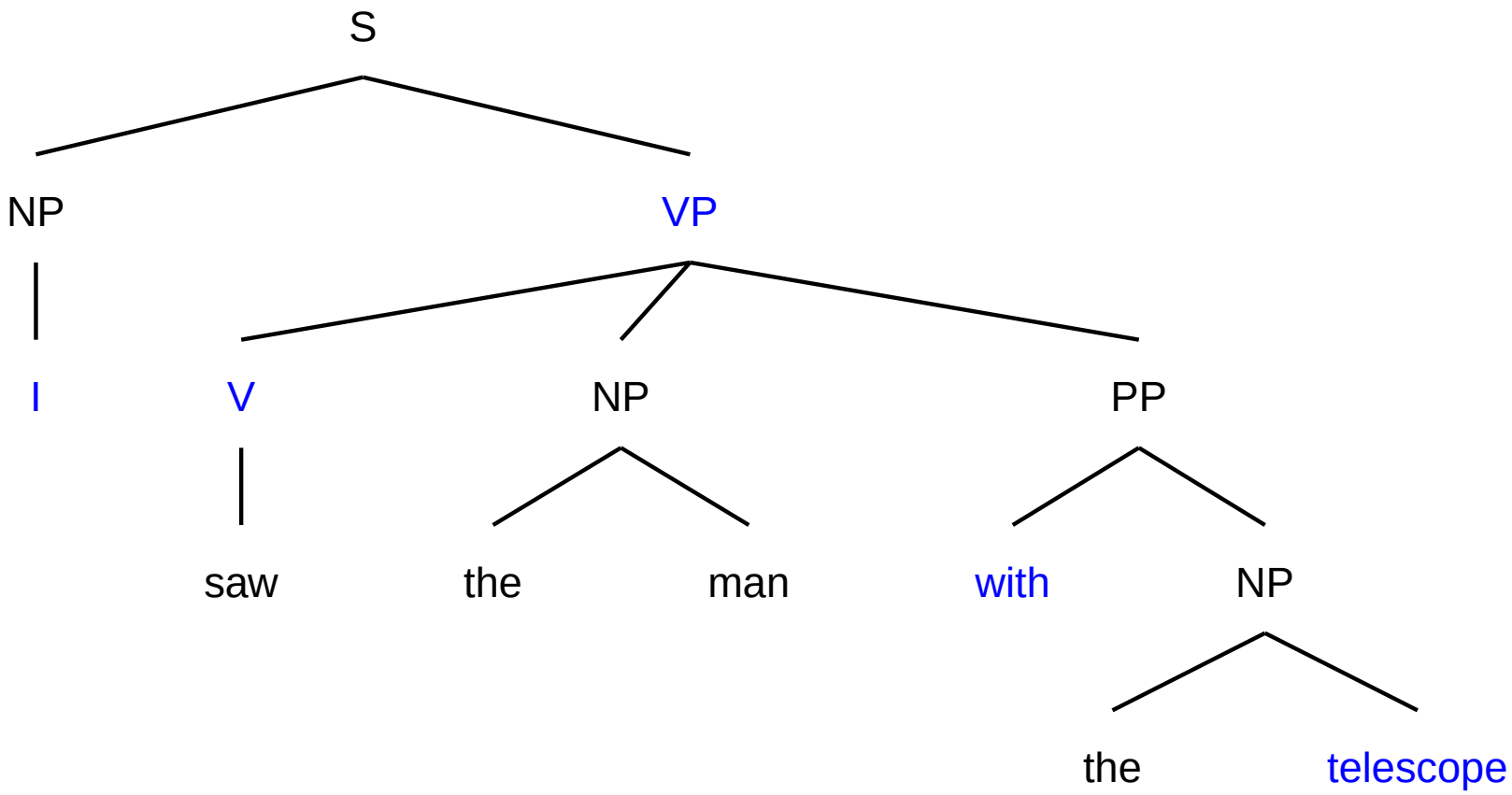


Stochastic Context Free Grammar

1	S $\Rightarrow$ NP VP
	...
0.5	VP $\Rightarrow$ V
0.25	VP $\Rightarrow$ V NP
0.25	VP $\Rightarrow$ V NP PP
	...
.0009	N $\Rightarrow$ *man*
.0001	V $\Rightarrow$ *man*

Independence assumption on trees.

Lexicalization

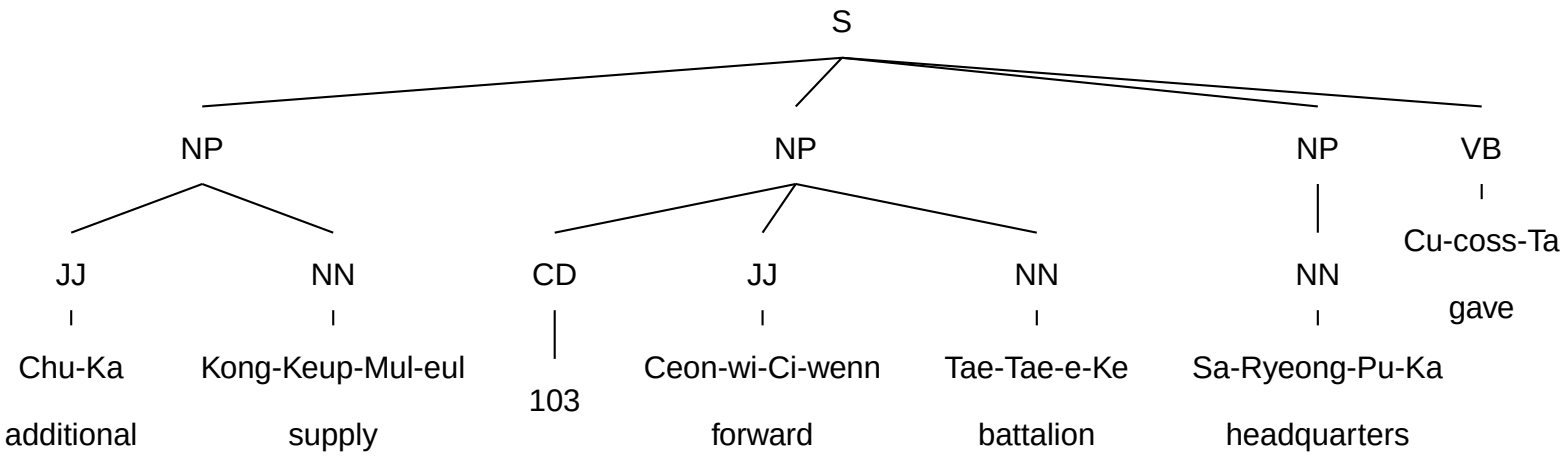


$$P(T) = P(\text{NP VP}|\text{S,saw})P(\text{V NP PP}|\text{VP,saw})P(\text{P NP}|\text{PP,with})\ldots$$

Translation Example

Korean	Chu-Ka Kong-Keup-Mul-eul 103 Ceon-wi-Ci-wenn- Tae-Tae-e-Ke Sa-Ryeong-Pu-Ka Cu-coss-Ta
Word gloss	Additional supply 103 forward support battalion headquarters gave
Commercial MT	Additional supply 103 FSB headquarters which you bite gave
Target	Headquarters gave the 103rd forward support battalion additional supplies

Syntactic Tree Representation



Meaning is More Than Syntactic Structure

- HQ gave [$_{\text{object}}$ the battalion] supplies

  HQ gave supplies [$_{\text{PP}}$ to the battalion]

- She broke [$_{\text{object}}$ the cup]

  [$_{\text{subject}}$ The cup] broke

- The discussion [$_{\text{PP}}$ between the leaders]

  [$_{\text{possessive}}$ The leaders'] discussion

Meaning requires argument structure, i.e., semantic roles.

Predicate Argument Representation

[$_{\text{Donor}}$ Headquarters] gave [$_{\text{Recipient}}$ the 103rd forward support battalion]

[$_{\text{Theme}}$ additional supplies]

Approaches to Natural Language Understanding

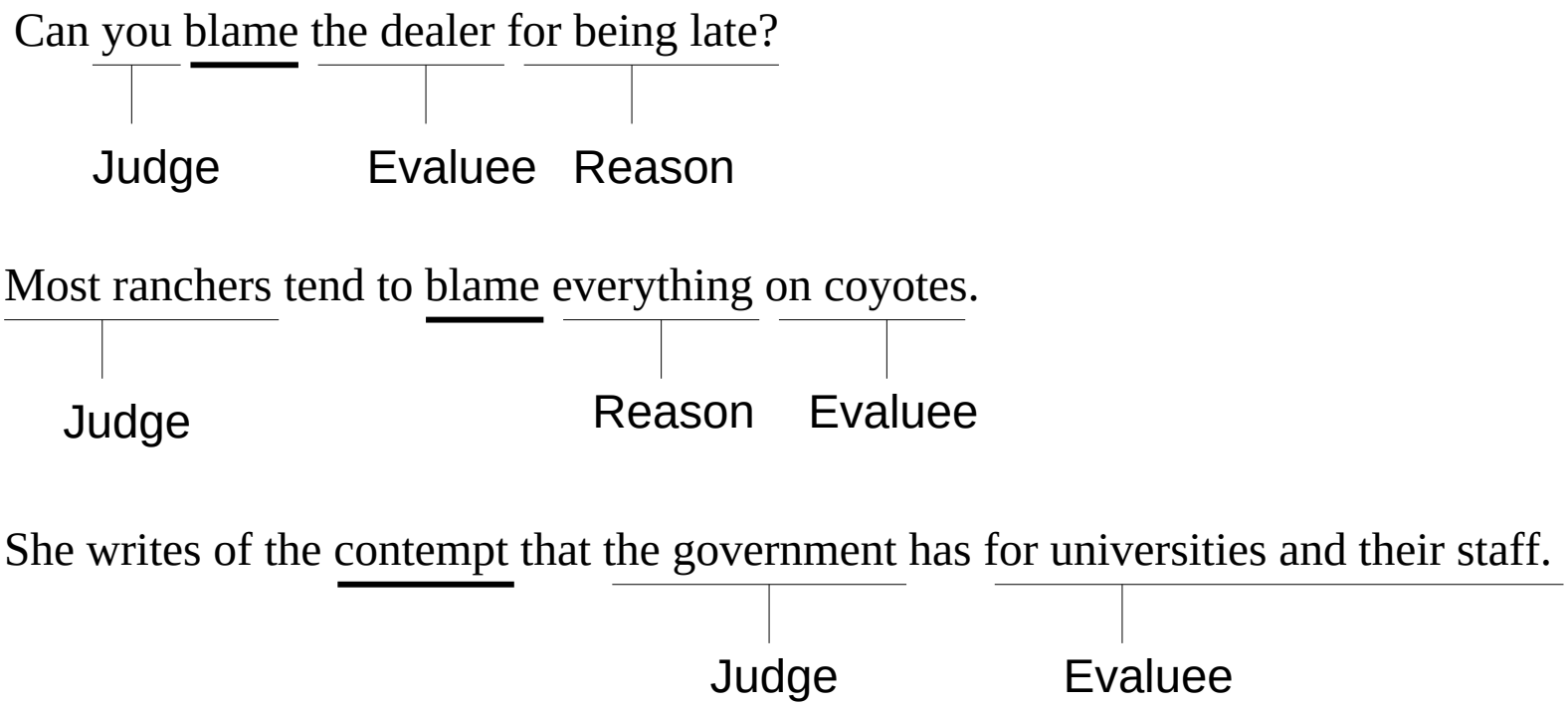
- Domain knowledge known to be important since Winograd (1972).
- Work during 1980s focused on deeper semantics: analysis of quantifiers, pronoun resolution, discourse structure.
  - fragile!
- Recent shift towards “shallow” semantics, systems trained on annotated data
- But these operate within a very constrained domain (MUC evaluations: corporate acquisitions, terrorist events)
- This talk: statistical system for general domains.

Current Directions in NLP

Statistical approaches to broad coverage semantics are beginning to be possible thanks to

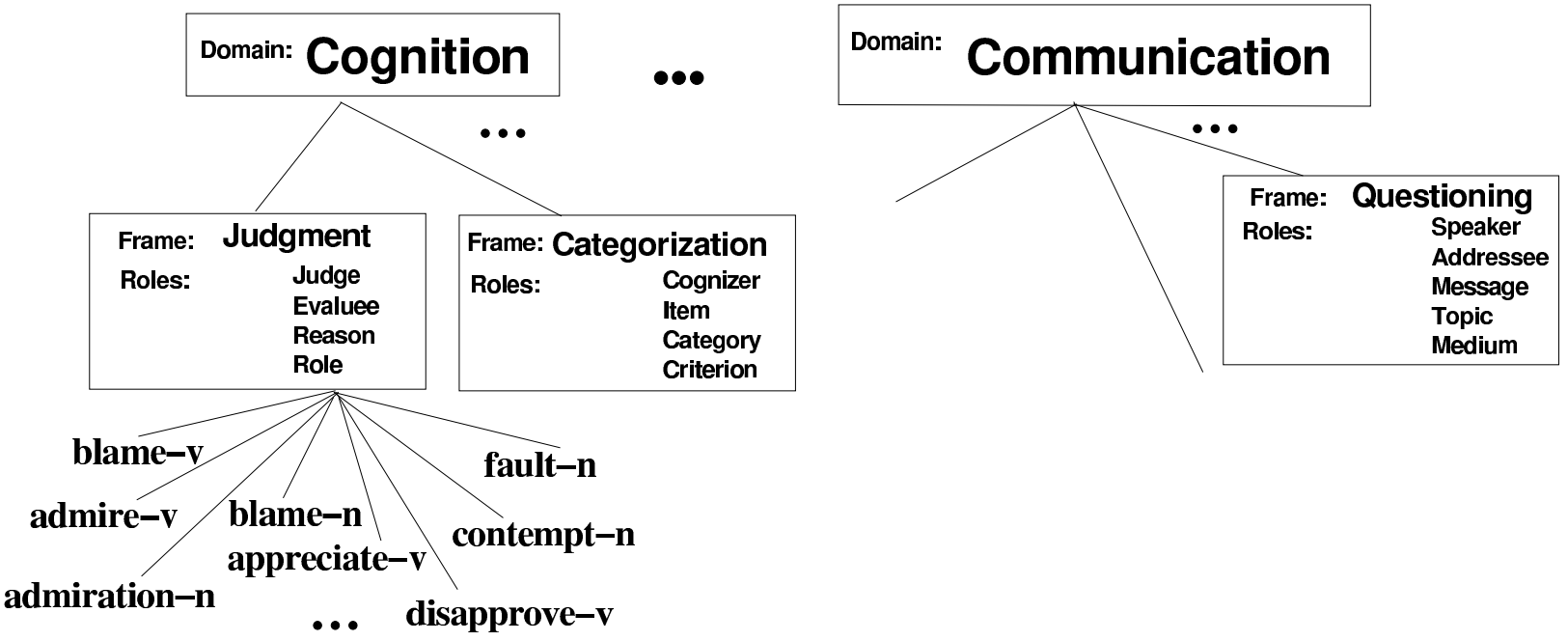
- Improvements in robust, statistical syntactic parsers
- Large amounts of text available (including parallel text)
- New learning techniques to leverage unlabeled data

Goal: Automatically Label Sentence with Semantic Roles



Semantic Frames – FrameNet

Frame Level:



The FrameNet Corpus

- 49,000 instances of predicates from British National Corpus, with 99,000 annotated arguments
- 1462 predicate words from 67 frames:
  - 927 verbs, 339 nouns, 175 adjectives

Proposition Bank

Kingsbury & Palmer (2002)

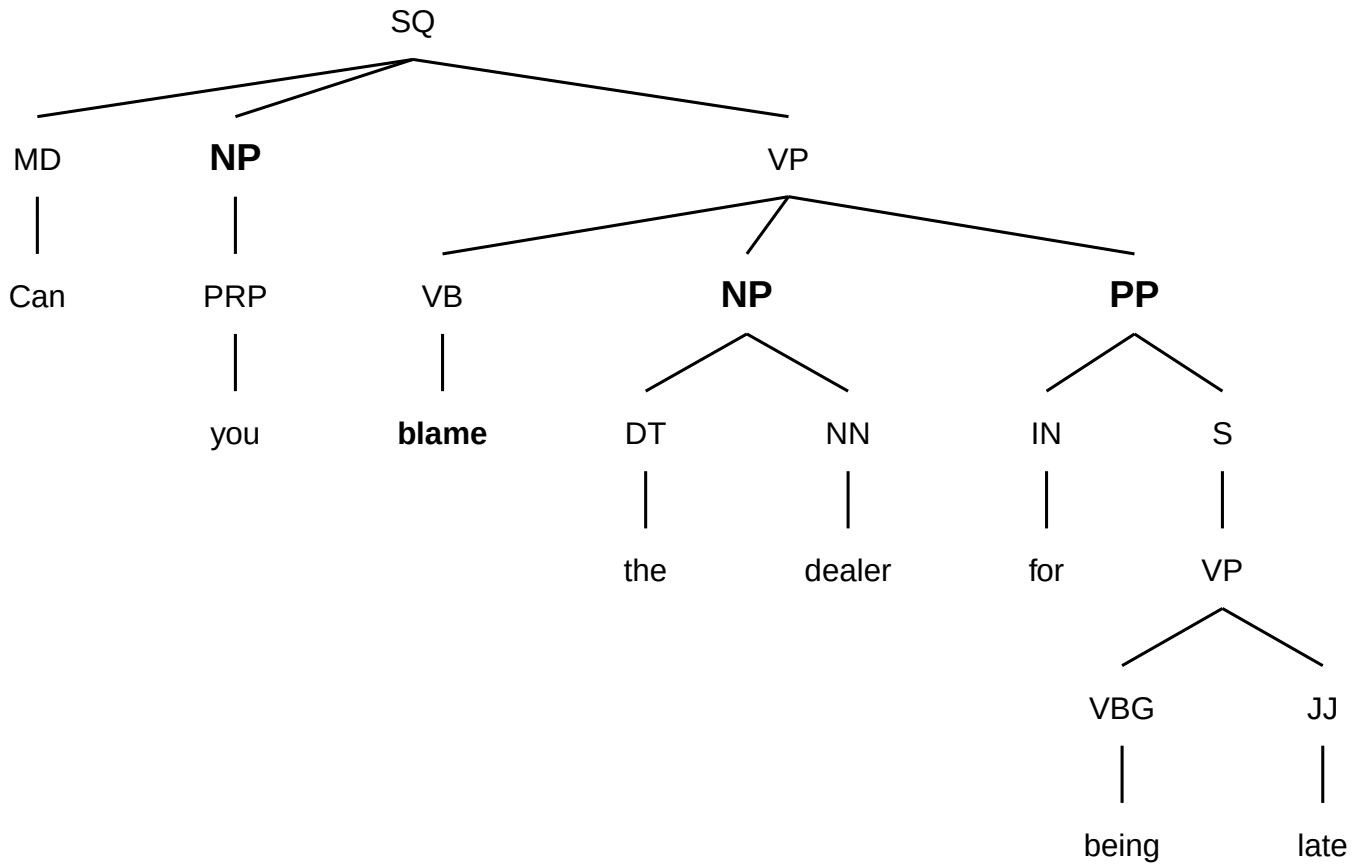
- Argument labels defined per-predicate: Arg0, Arg1, ... Temporal, Locative, etc
- Tagging all verbs in Wall Street Journal corpus, for which syntactic trees are available
- Preliminary corpus: 26,000 instances of predicates, 65,000 individual arguments, 1527 unique verbs

Probability Model for Roles

- Extract set of features F_i for each constituent i in syntactic tree for sentence containing predicate p
- Role probabilities for individual constituents: $P(r_i|F_i, p)$
- Probabilities of predicate's complete set of roles: $P(r_{1..n}|F_{1..n}, p)$

Feature 1: Phrase Type

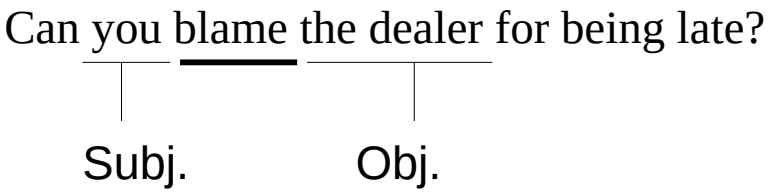
Can [$_{Judge}$ you] **blame** [$_{Evaluee}$ the dealer] [$_{Cause}$ for being late] ?



From output of automatic parser (Collins)

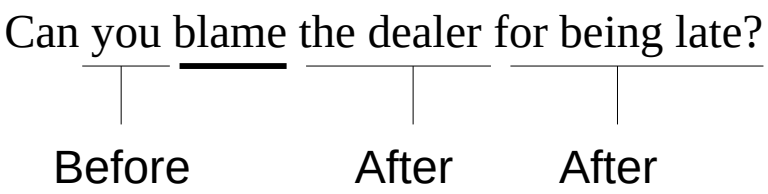
Feature 2: Grammatical Function

Read from parse tree, used for NP constituents only:



Feature 3: Position

Whether constituent is Before/After predicate word:



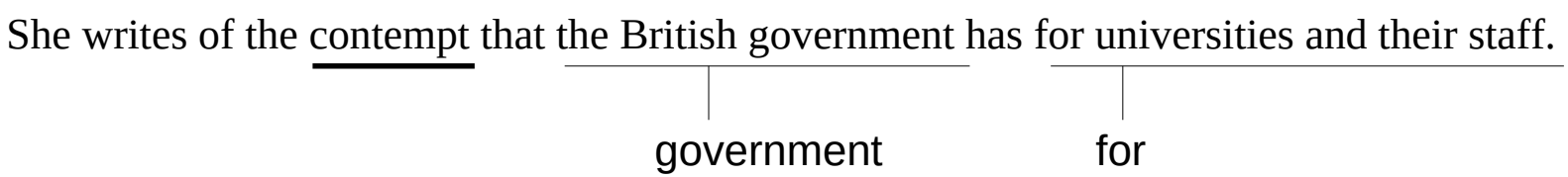
Feature 4: Voice

Active/Passive use of predicate word read from parse tree:

"Can you **blame** the dealer" vs. "The dealer **is blamed**"

Feature 5: Head Word

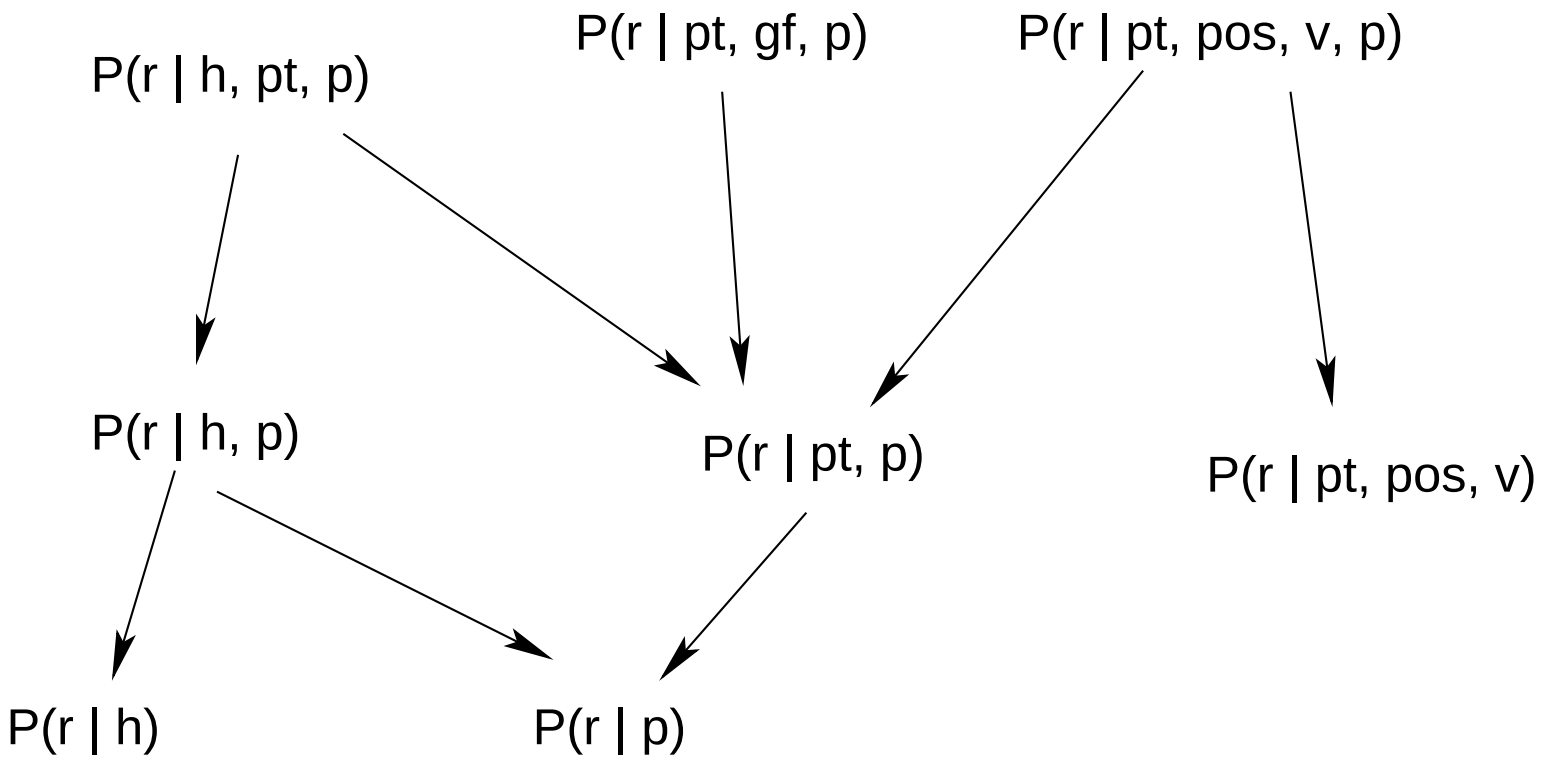
Head word of constituent as defined by parser:



Probability Model for Roles

- Extract features $F = \{pt, gf, pos, v, h\}$ from syntactic parse
- Role probabilities for individual constituents: $P(r|pt, gf, pos, v, h, p)$
- Sparseness of training data prevents direct estimation
- Combine probabilities conditioned on subsets of features

Backoff Lattice



Combining Distributions

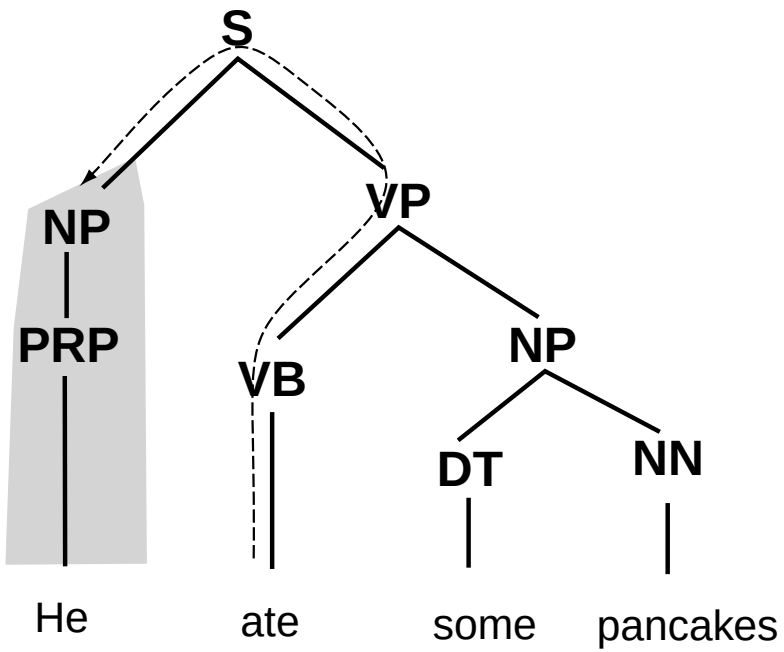
Distributions	*Combining Method*	*Correct*
Baseline	$P(r\|p)$	40.9%
All	Equal Linear Interpolation	79.5
	Weighted Linear Interpolation	79.3
	Geometric Mean	79.6
Backoff	Linear interpolation	80.4
	Geometric mean	79.6

% labeled with correct role, 8167 observations

Finding Which Constituents Are Arguments

Calculate probabilities of a constituent being an argument given features:

- Path through tree from predicate to constituent
- Predicate
- Head Word



Ex: $P(arg = \text{true} \mid p = \text{“eat”}, path = \text{“VB}\uparrow\text{VP}\uparrow\text{S}\downarrow\text{NP”}, head = \text{“He”})$

Probabilities for Sets of Arguments

Probability $P(\{r_{1..n}\}|p)$ of unordered set of arguments observed in a sentence:

$\{$ JUDGE, EVALUEE, REASON $\}$

"Can you blame the dealer for being late?"

"Ranchers tend to blame everything on coyotes."

Argument Set	$P(\{r_{1..n}\}\|p =$"blame"$)$
$\{$ EVAL, JUDGE, REAS $\}$	0.549
$\{$ EVAL, JUDGE $\}$	0.160
$\{$ EVAL, REAS $\}$	0.167
$\{$ EVAL $\}$	0.097
$\{$ EVAL, JUDGE, ROLE $\}$	0.014
$\{$ JUDGE $\}$	0.007
$\{$ JUDGE, REAS $\}$	0.007

Interdependence of Argument Assignments

Choose best assignment of roles $r_{1..n}$ given predicate p, and features $F_{1..n}$:

$$
\begin{aligned}
argmax_{r_{1..n}} P(r_{1..n}|F_{1..n},p) &= P(r_{1..n}|p)\frac{P(F_{1..n}|r_{1..n},p)}{P(F_{1..n}|p)} \\
&\approx P(\{r_{1..n}\}|p)\prod_i P(F_i|r_i,p) \\
&= P(\{r_{1..n}\}|p)\prod_i \frac{P(r_i|F_i,p)P(F_i|p)}{P(r_i|p)} \\
&= P(\{r_{1..n}\}|p)\prod_i \frac{P(r_i|F_i,p)}{P(r_i|p)}
\end{aligned}
$$

Argument set probabilities provide (limited) dependence between individual labeling decisions.

Integrated Labeling and Boundary Identification

	Prec.	Recall
Label constituents independently: $\prod_i P(r_i \mid F_i, p)$	67.0	46.8
With argument sets: $P(\{r_{1..n}\}\mid p) \prod_i \frac{P(r_i \mid F_i, p)}{P(r_i \mid p)}$	64.6	61.2

Argument sets improve recall by telling the system what it's looking for.

Effect of Parsing Accuracy

PropBank data come with gold-standard syntactic trees.

	FrameNet		*PropBank*	
	Precision	*Recall*	*Precision*	*Recall*
Gold-standard parses			73.5	71.7
Automatic parses	64.6	61.2	59.0	55.4

Accurate syntactic parsing is important!

But even automatically generated trees are better than a flat representation.

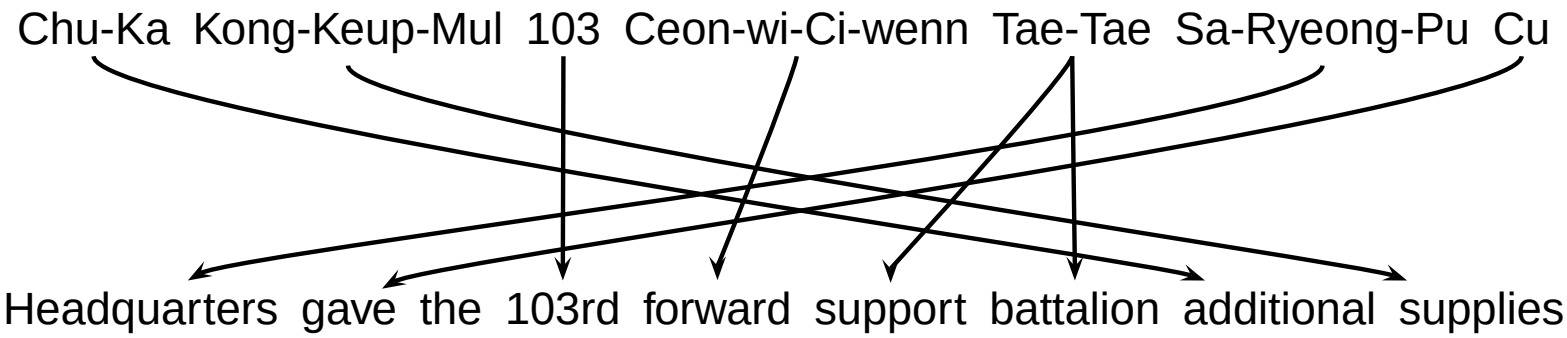
Machine Translation

Current approaches:

- Analyze source language, generate in target language (commercial systems)
- Statistical approaches trained on parallel text
  - Two probability models estimated from parallel corpora:
    word translation and word-level alignment.

Alignment Example

An alignment is a set of pairs of words which correspond:

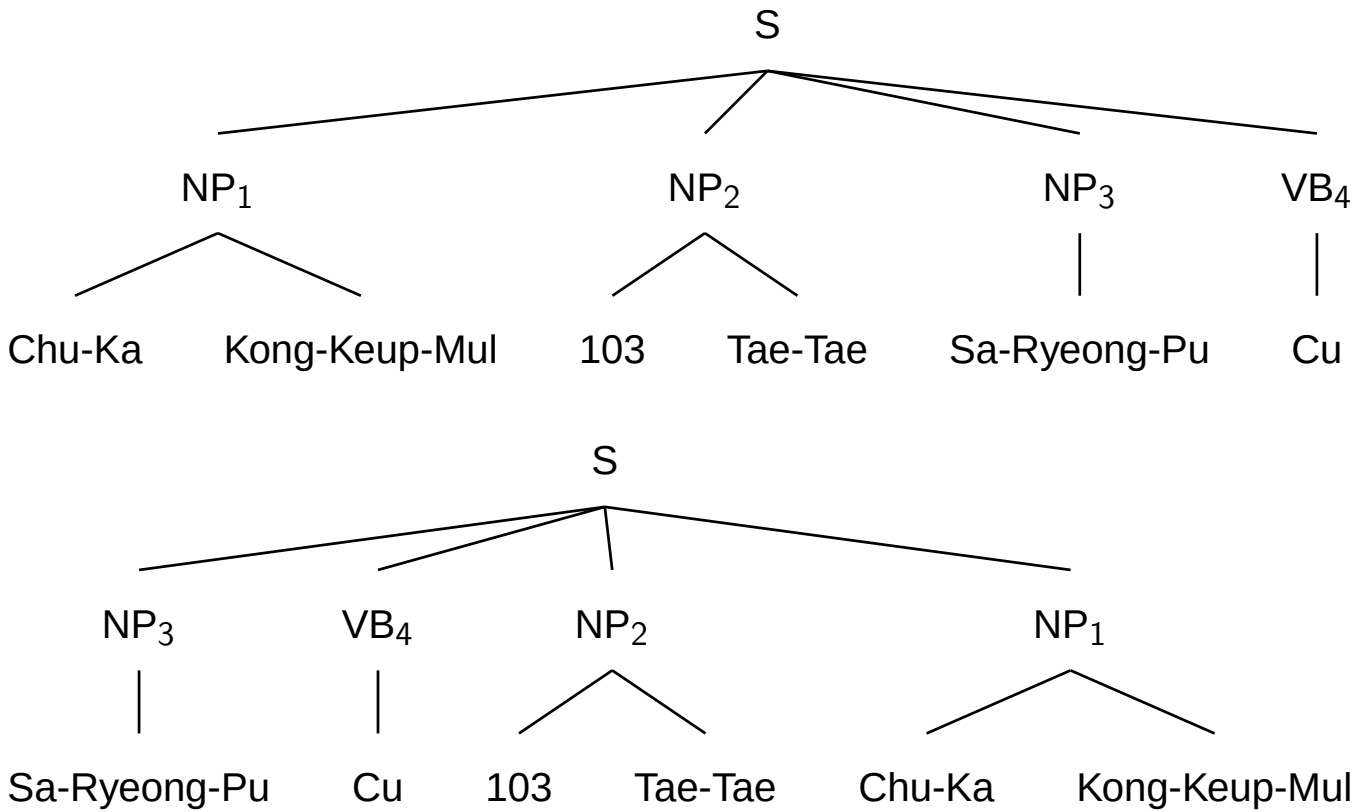


A translation probability: $P_t(\text{Headquarters} \mid \text{Chu-Ka})$

An alignment probability: $P_a(e_8|k_1)$

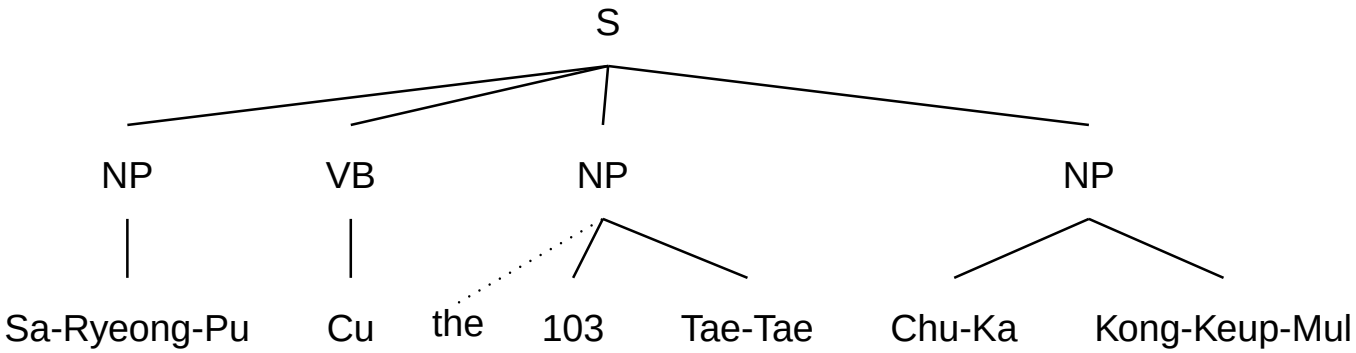
Tree-Based Alignment

Yamada & Knight 2001

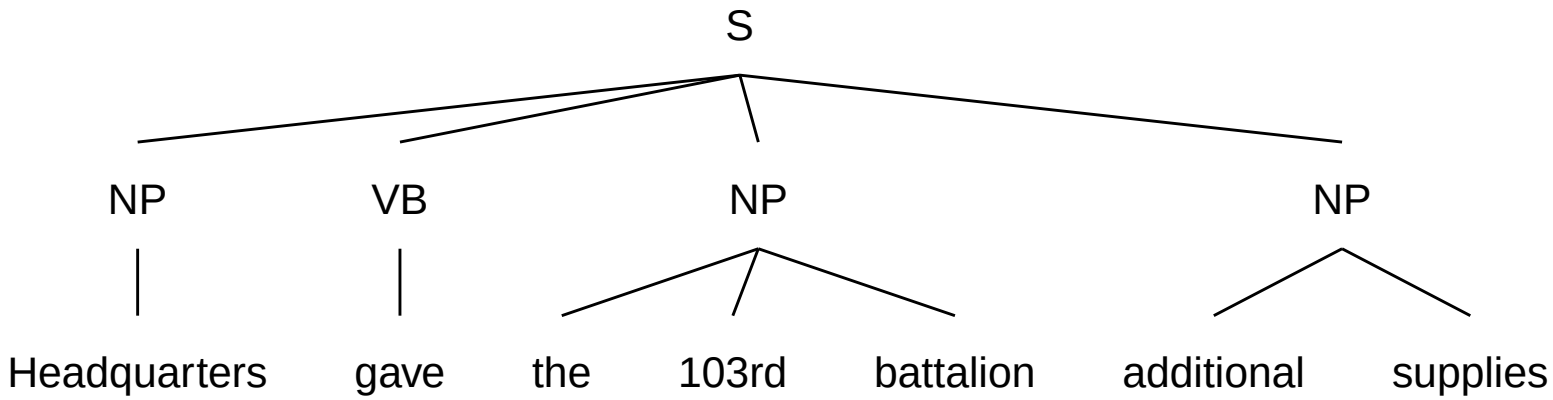


re-order step: $P_r(3, 4, 2, 1 \mid \mathrm{S} \Rightarrow \mathrm{NP\ NP\ NP\ VB})$

Tree-Based Alignment 2



insertion step: $P_{ins}(\text{the})P(ins|NP)$

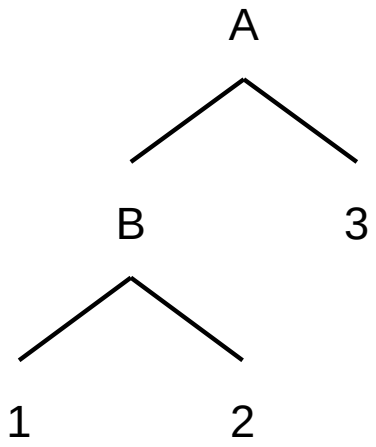


translation step: $P_t(\text{give}|\text{Cu})$

EM Training Procedure

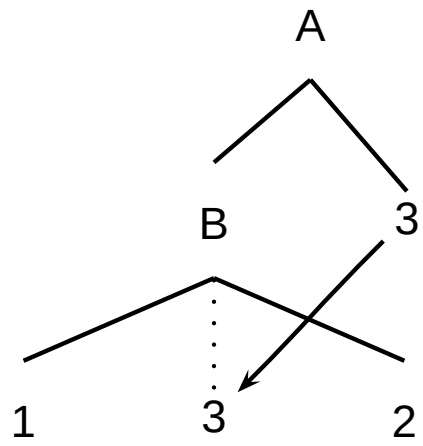
- Observed data: sentence pairs
- Hidden data: word-level alignment
- Compute expected counts for each possible alignment using dynamic programming (E-step)
- Re-estimate re-order, insert, and translation probabilities (M-step)

Trees Constrain Possible Alignments



Of the six possible re-orderings of the three terminals, two are not allowed: 1,3,2 and 2,3,1

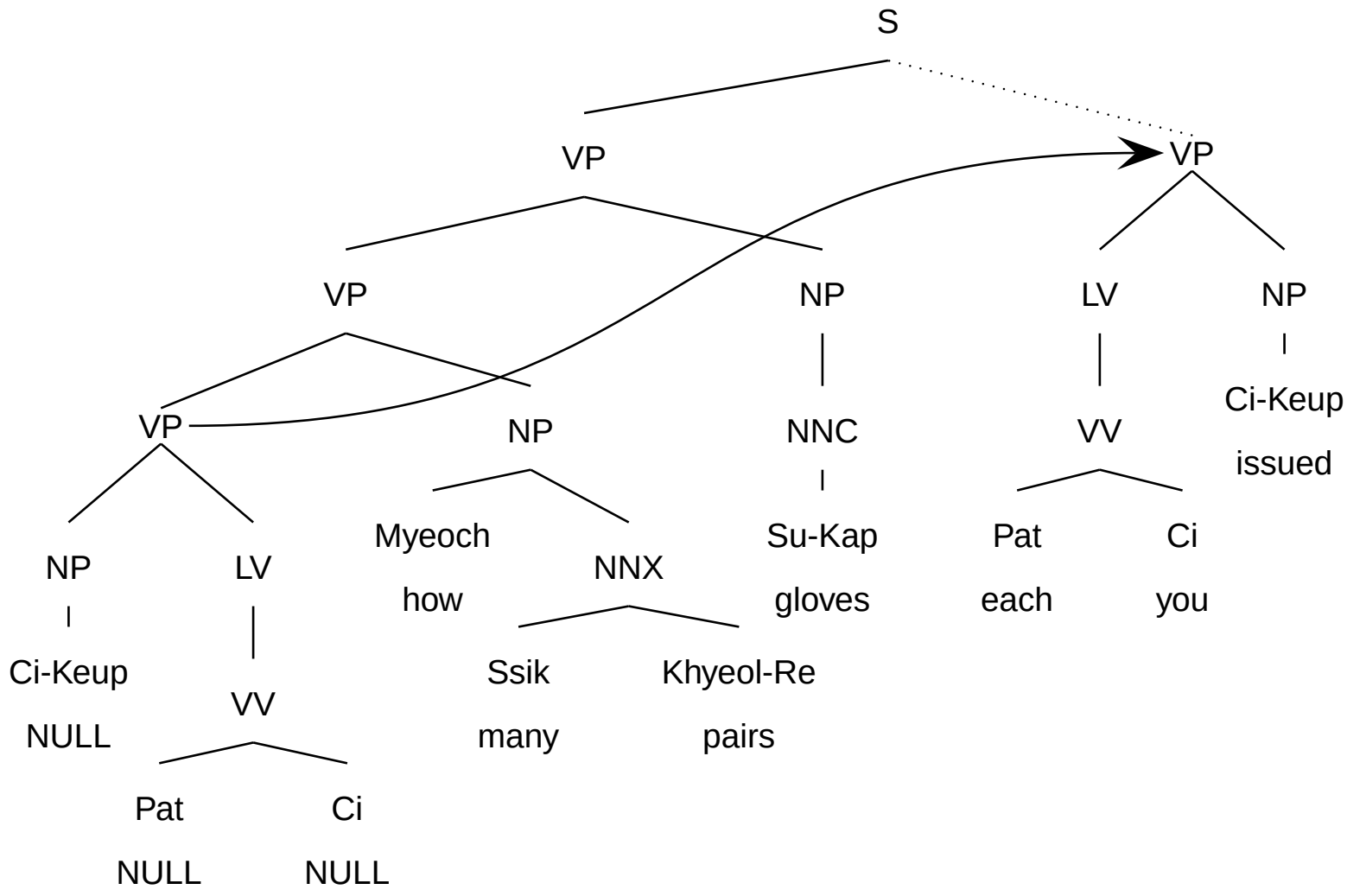
Allow Subtrees to be “Cloned”



Constituents of sentence can move to arbitrary locations, at a cost in probability.

Assumption that individual clone operations are independent means no increase in computational complexity.

Cloning example



English question word order not possible through tree re-ordering.

Korean-English Parallel Treebank

- 5083 parallel sentences
- Human annotated syntactic trees in both languages
- Average English sentence is 16 words

Results: Alignment Error

Agreement with human judgments of word-level alignment.

	Alignment Error Rate
Unstructured (IBM)	.35
Tree Re-order (Y&K)	.43
Tree Re-order, Clone	.32

- Tree-based model is **too rigid** by itself.
- Clone operation provides flexibility needed to use syntactic information in statistical system.

Alignment - Summary

Relaxing tree-based model improves alignments, hybrid between purely statistical and analytic systems, and combines the robustness of statistical methods with the benefits of syntactic analysis.

Conclusion

- General approach to ambiguity: Learn from examples
- Key Problems:
  - sparse data
  - generalization
  - providing the right representation