Foundations of Statistical Machine Translation: Past, Present and Future

Taro Watanabe
taro.watanabe @ nict.go.jp
http://mastarpj.nict.go.jp/~t_watana/

20 years history

- Statistical Machine Translation (SMT) started from Brown et al. (1990)
- Is SMT matured?
  - Real service: Web-based (Google, Microsoft), mobile phone (NICT)
- Promising gains from Tree-based approaches
  - Syntax-based SMT in {tree, string}-to-{tree, string}
  - Decoding = Parsing
- Better model, better search and better training

Statistical Machine Translation?

- MT as a decision making process:
  - Given a source text, search for the best translation
- Difference from Rule-based (Knowledge-based) MT:
  - Learn model/parameters from data
- Difference from Example-based MT:
  - Both are empirical, but more emphasis on examples + (usually) greedy search + heuristics

Overview of overview

- Foundation
  - Model, Training, Decoding
  - Phrase-based SMT
- Tree-based SMT
- Advanced Topics

Foundation

Overview

- Model, Training, Decoding
- Word Alignment
- Phrase-based SMT
- Evaluation
- Optimization

Translation as a decision problem

- Modeling:
  - Good p(e|f) approximating Pr(e|f)
  - Linguistic clues will be helpful
- Training:
  - Assign parameters given data
  - Maximum-likelihood, EM-algorithms, Bayesian
- Search:
  - Find the best translation
  - DP-based search with heuristic pruning

Source Channel Model

$$
\begin{aligned}
\hat{\mathbf{e}} &= \operatorname*{argmax}_{\mathbf{e}} Pr(\mathbf{e}|\mathbf{f}) \\
&= \operatorname*{argmax}_{\mathbf{e}} \frac{Pr(\mathbf{f}|\mathbf{e})Pr(\mathbf{e})}{Pr(\mathbf{f})} \\
&= \operatorname*{argmax}_{\mathbf{e}} Pr(\mathbf{f}|\mathbf{e})Pr(\mathbf{e}) \\
&= \operatorname*{argmax}_{\mathbf{e}} p(\mathbf{f}|\mathbf{e})p(\mathbf{e})
\end{aligned}
$$

- Early statistical machine translation (Brown et al., 1990)
- Since we do not know true distribution, we will approximate Pr(f|e) by p(f|e)

Source Channel Model

- Translation Model: p(f|e)
  - Bilingual correspondence between two sentences, f and e
  - Usually encode linguistic clues, such as dictionary
- Language Model: p(e)
  - “fluency” for the generated sentence

Log-linear Model

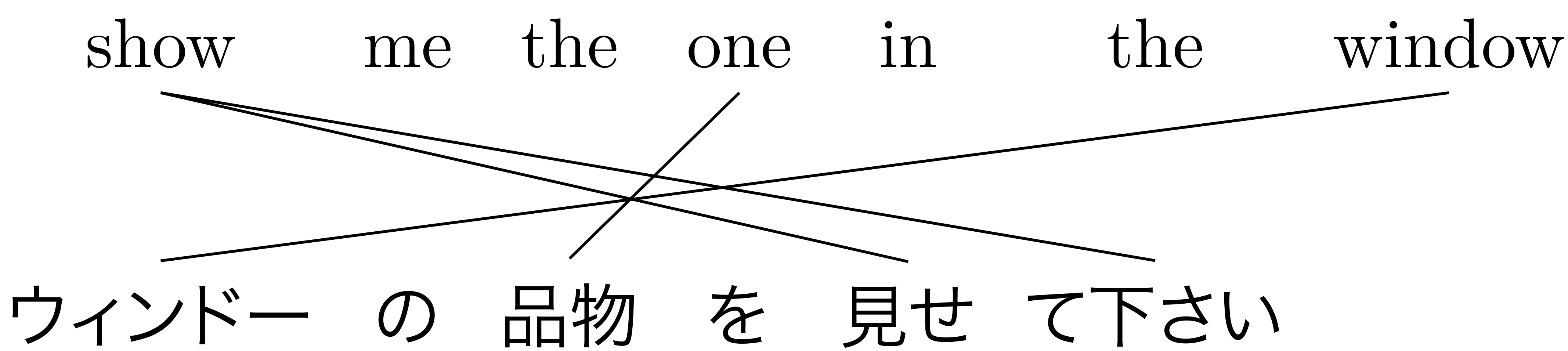
$$p(\mathbf{e}|\mathbf{f}) = \frac{\exp\left(\mathbf{w} \cdot \mathbf{h}(\mathbf{e}, \mathbf{f})\right)}{\sum_{\mathbf{e}'} \exp\left(\mathbf{w} \cdot \mathbf{h}(\mathbf{e}', \mathbf{f})\right)}$$

- Generalization of Source Channel model
- Each feature function captures different aspect of translations
- Each feature function is weighted
- Easy to incorporate new features

Overview

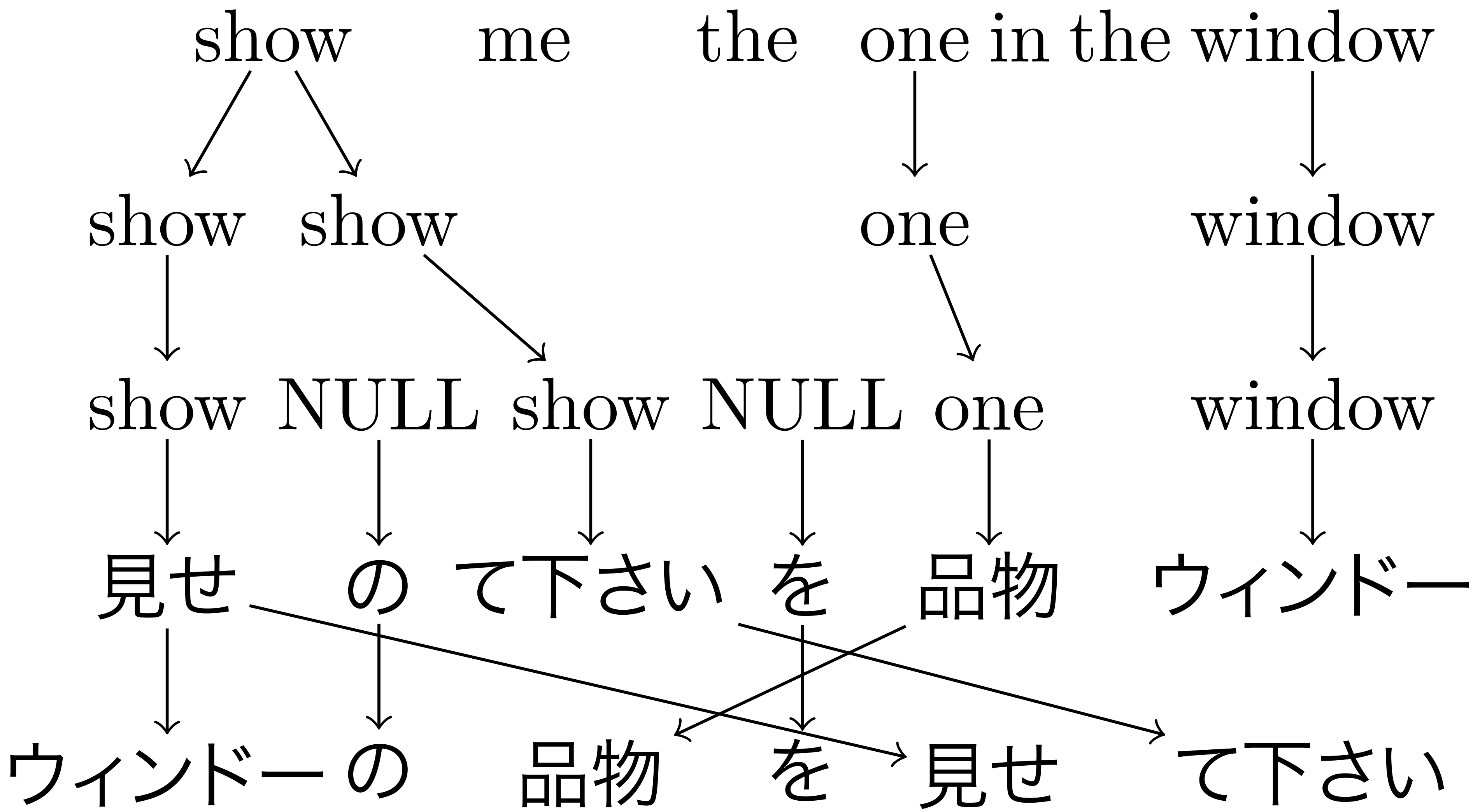
- Model, Training, Decoding
- Word Alignment
- Phrase-based SMT
- Evaluation
- Optimization

Word alignment



- One of the fundamental unit of translation
  - one-to-one correspondence
  - or, many-to-many alignment

Word alignment models



- IBM Model 4
- Decompose into several models: fertility, lexicon, distortion

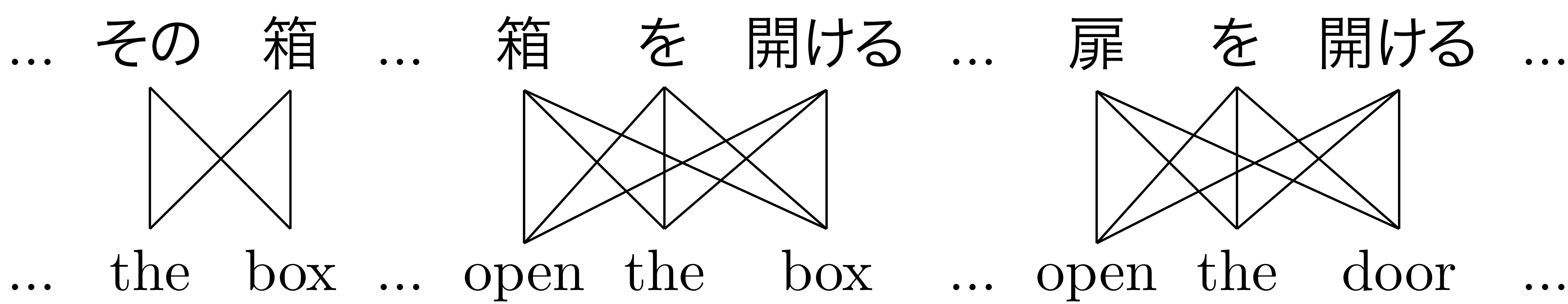
Word alignment models

$$p(\mathbf{a}, \mathbf{f}|\mathbf{e}) = \sum_{j=1}^{J} p_d(\mathbf{a}_j|\mathbf{a}_{j_-}, j) p_t(\mathbf{f}_j|\mathbf{e}_{\mathbf{a}_j})$$

$$\begin{aligned}
p_d(\mathbf{a}_j = 0|\mathbf{a}_{j_-} = i) &= p_0 \\
p_d(\mathbf{a}_j = i' \neq 0|\mathbf{a}_{j_-} = i) &\propto (1 - p_0) \begin{cases} 1 & \text{(IBM 1)} \\ c(i' - \lfloor \frac{jI}{J} \rfloor) & \text{(IBM 2)} \\ c(i' - i) & \text{(HMM)} \end{cases}
\end{aligned}$$

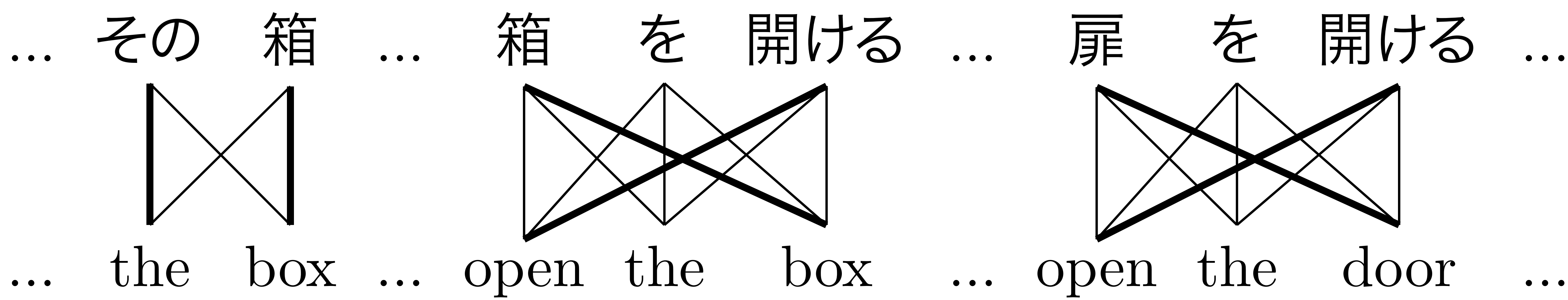
- IBM 1, IBM 2 and HMM
- More models, such as IBM {3,4,5}

Word alignment training



- EM algorithm:
  - E-step to compute expected counts
  - M-step to perform maximization

Word alignment training



- Starting from uniform parameter, try compute expectation of aligning words
- Based on the expectation, estimate parameters
- Iterate....until convergence

Word alignment model training

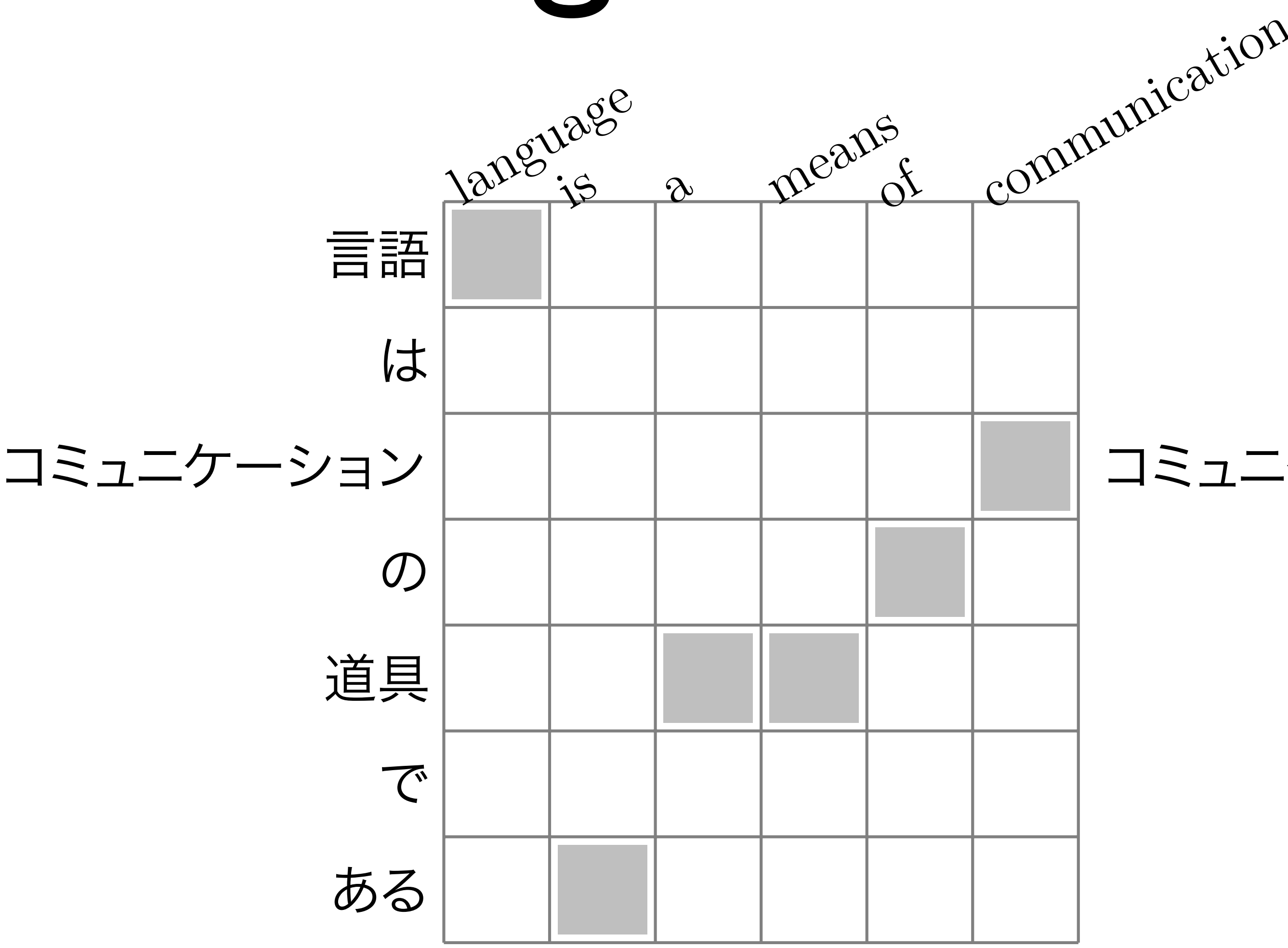
$$\begin{aligned}\hat{\theta} &= \operatorname*{argmax}_{\theta} \prod_{\mathbf{e},\mathbf{f}} p(\mathbf{f}, \mathbf{a}|\mathbf{e}; \theta) \\ &= \operatorname*{argmax}_{\theta} \sum_{\mathbf{e},\mathbf{f}} \log p(\mathbf{f}, \mathbf{a}|\mathbf{e}; \theta)\end{aligned}$$

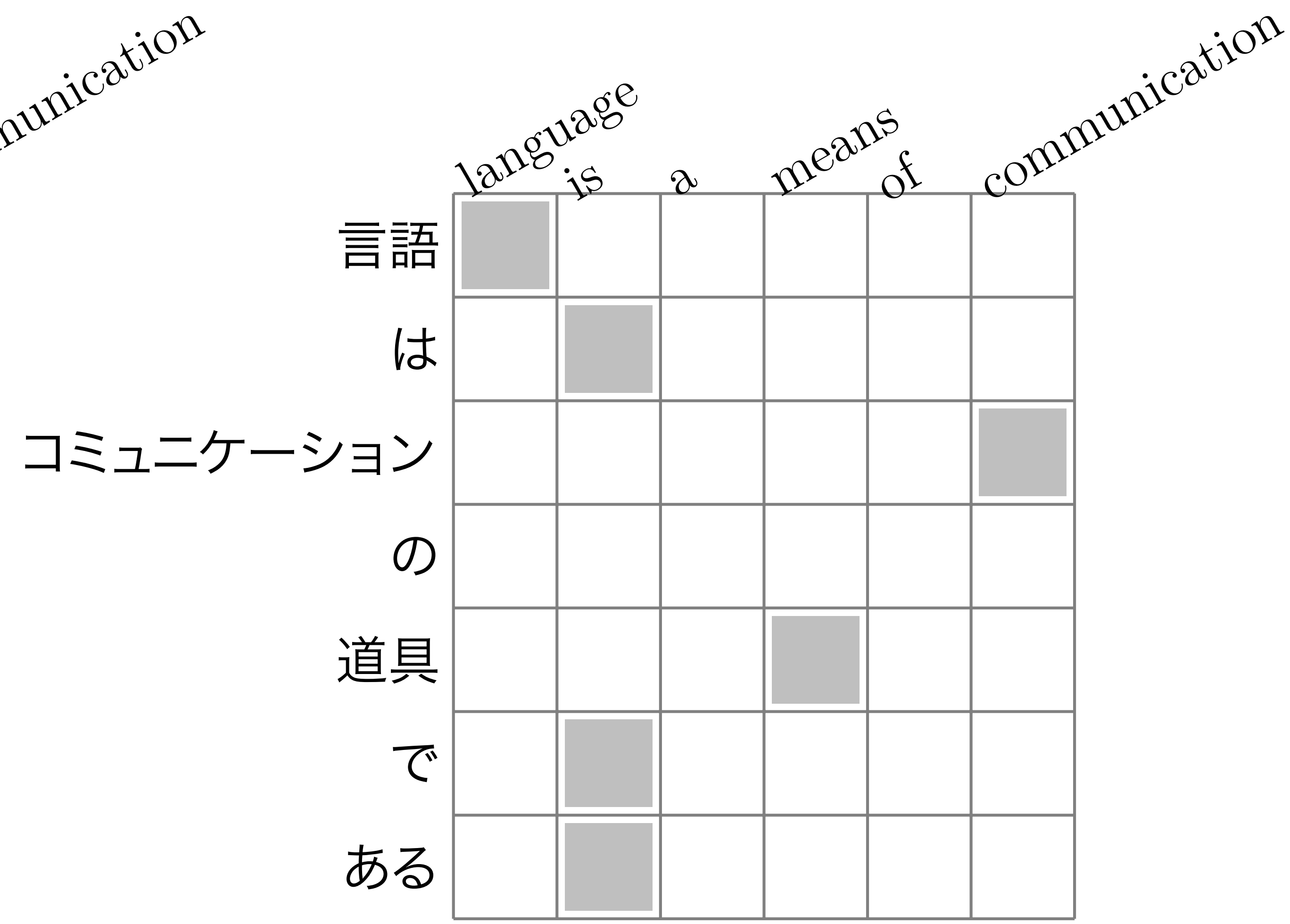
E-step: $q(\mathbf{a}; \mathbf{f}, \mathbf{e}) = p(\mathbf{a}|\mathbf{e}, \mathbf{f}; \theta)$

M-step: $\theta' = \operatorname*{argmax}_{\theta} \sum_{\mathbf{f},\mathbf{e},\mathbf{a}} q(\mathbf{a}; \mathbf{f}, \mathbf{e}) \log p(\mathbf{f}, \mathbf{e}, \mathbf{a}; \theta)$

- Inside EM-training
    - Maximizing log-likelihood over the training data

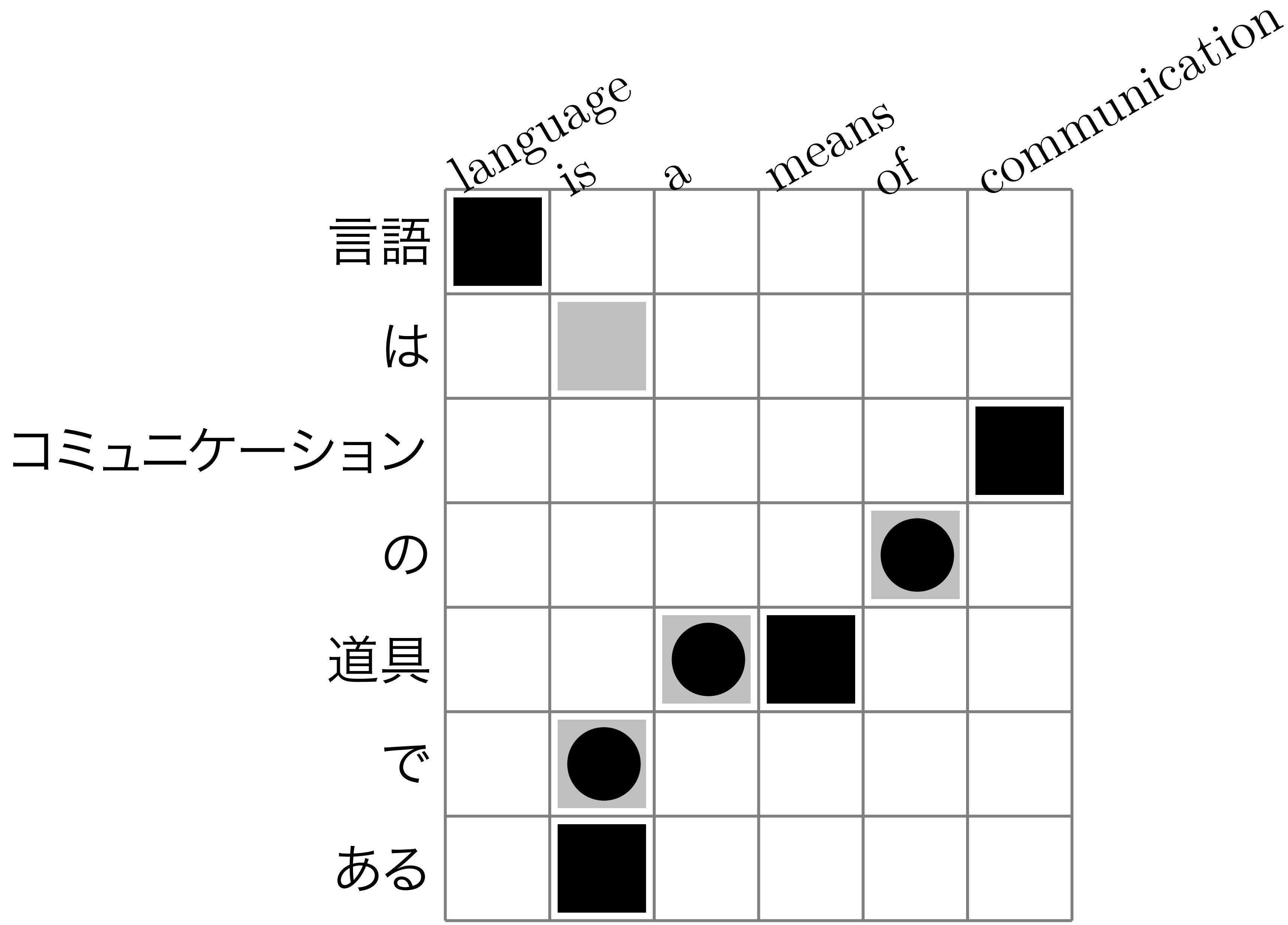
Alignment combination





- IBM Models are limited to one-to-many
- Prone to errors, especially for rare words
- Training in both directions, "heuristically" combine

Alignment heuristics



- Starts from intersected alignment, greedily add union alignments

Symmetric training

$$\text{E-step:} \quad q(\mathbf{a};\mathbf{f},\mathbf{e}) = \frac{1}{Z_{\mathbf{f},\mathbf{e}}} p_1(\mathbf{a}|\mathbf{f},\mathbf{e};\theta_1) \cdot p_2(\mathbf{a}|\mathbf{e},\mathbf{f};\theta_2)$$

$$\text{M-step:} \quad \theta' = \operatorname*{argmax}_{\theta} \sum_{\mathbf{f},\mathbf{e},\mathbf{a}} q(\mathbf{a};\mathbf{f},\mathbf{e}) \log p_1(\mathbf{f},\mathbf{e},\mathbf{a};\theta_1) + \sum_{\mathbf{f},\mathbf{e},\mathbf{a}} q(\mathbf{a};\mathbf{f},\mathbf{e}) \log p_2(\mathbf{f},\mathbf{e},\mathbf{a};\theta_2)$$

(Liang et al., 2006)

- Alternatives to heuristic approaches, it is possible to approximate symmetization during EM-algorithm
  - Jointly maximize both directions by approximating summation (Liang et al., 2006)
  - Consider additional agreement constraint and minimize KL divergence (Ganchev et al., 2008)

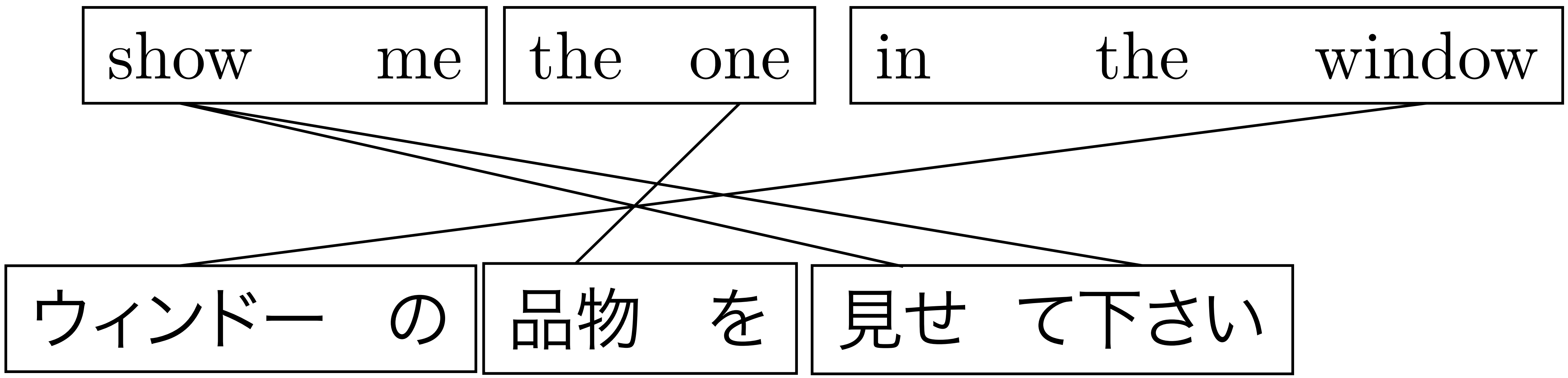
Decoder for word alignment models?

- Possible, but prone to errors
  - NP-hard problem (Knight, 1999)
  - Many alternative translations with insertion/deletion
  - Spurious reordering: no distinction with local/global reordering

Overview

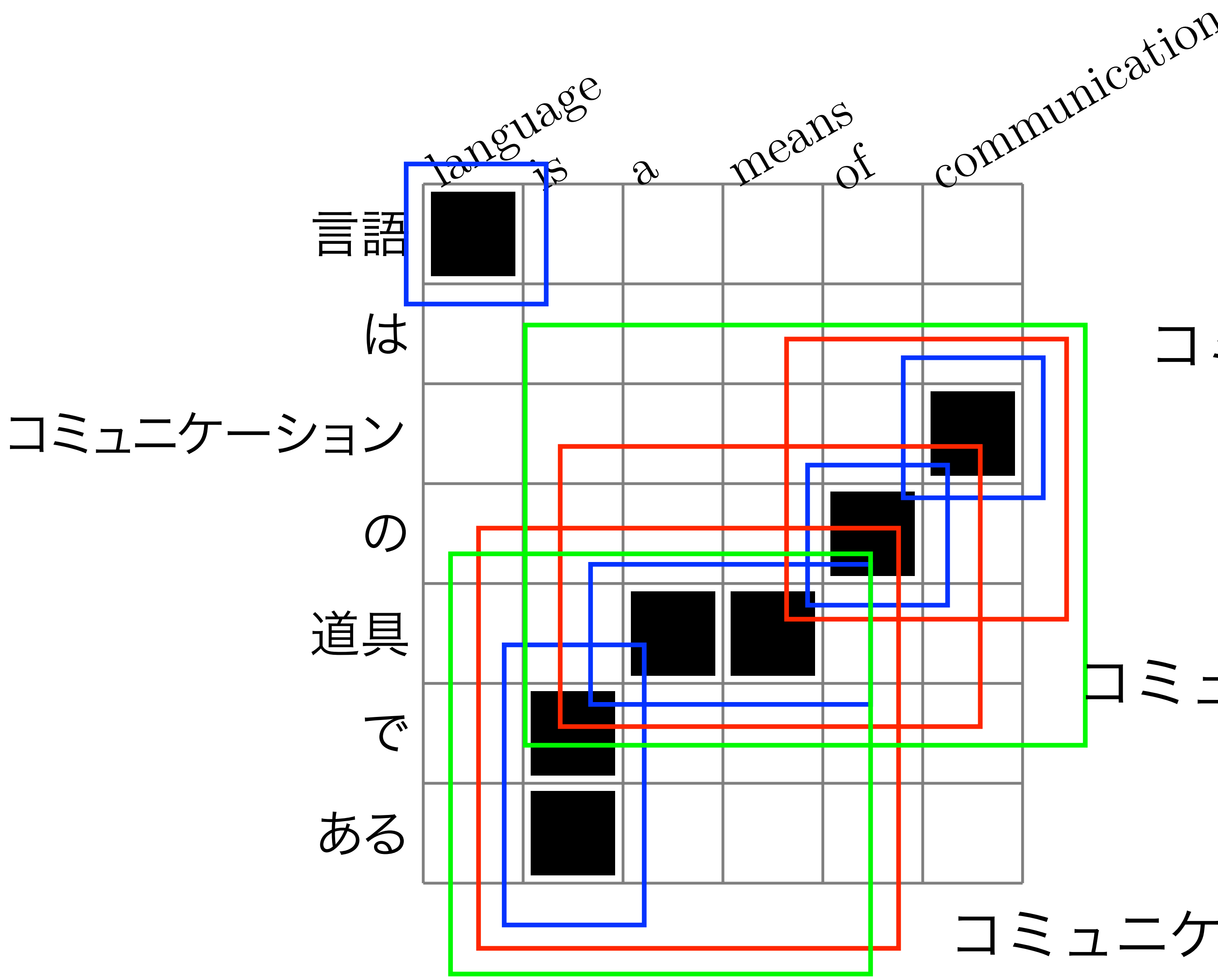
- Model, Training, Decoding
- Word Alignment
- Phrase-based SMT
- Evaluation
- Optimization

Phrase-based SMT



- Directly employing word-based model for decoding is not practical
  - Many decisions: local/global reordering, insertion/ deletion
- Use phrases to capture local reordering (at least)

Phrase extraction



言語	language
コミュニケーション	communication
の	of
道具	a means
で ある	is
コミュニケーション の	of communication
の 道具	a means of
道具 で ある	is a means of
コミュニケーション の 道具	a means of communication

- Given word alignment, contiguous phrases are extracted which do not violate alignment constraint
- Relative count-based estimation + smoothing

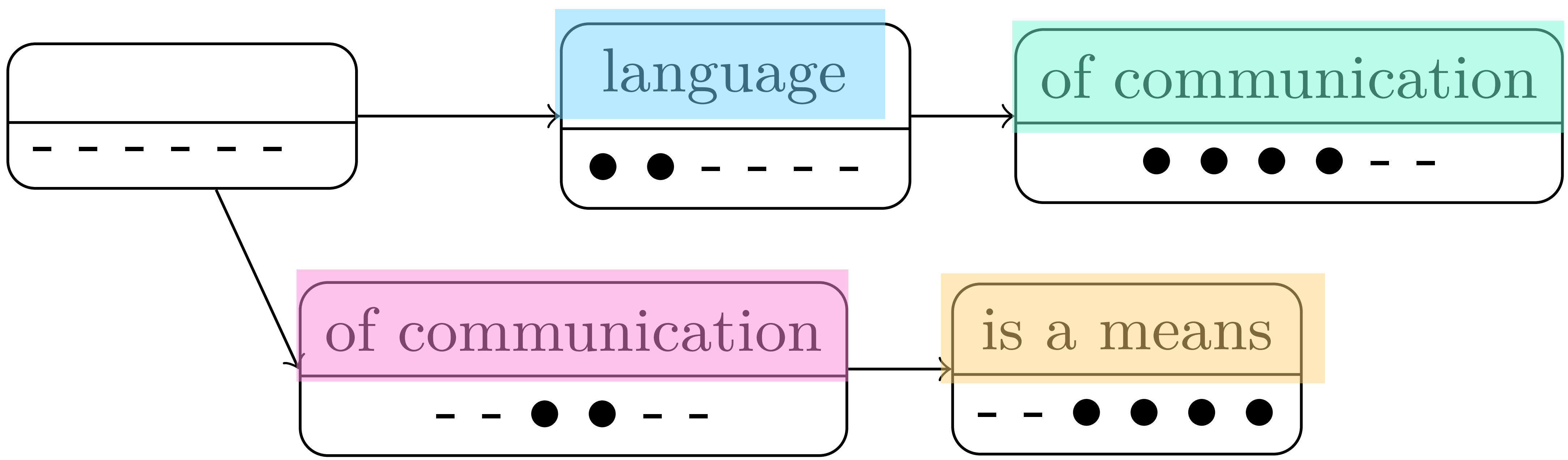
Decoding for phrase-based SMT

$$
\begin{aligned}
\hat{\mathbf{e}} &= \operatorname*{argmax}_{\mathbf{e}} \frac{\exp\left(\mathbf{w} \cdot \mathbf{h}(\mathbf{e}, \phi, \mathbf{f})\right)}{\sum_{\mathbf{e}', \phi'} \exp\left(\mathbf{w} \cdot \mathbf{h}(\mathbf{e}', \phi', \mathbf{f})\right)} \\
&= \operatorname*{argmax}_{\mathbf{e}} \mathbf{w} \cdot \mathbf{h}(\mathbf{e}, \phi, \mathbf{f})
\end{aligned}
$$

- Maximization by log-linear model with hidden phrase structures
- Φ: hidden variable for phrasal segmentation
- Max-derivation: searching for the best segmentation + translation

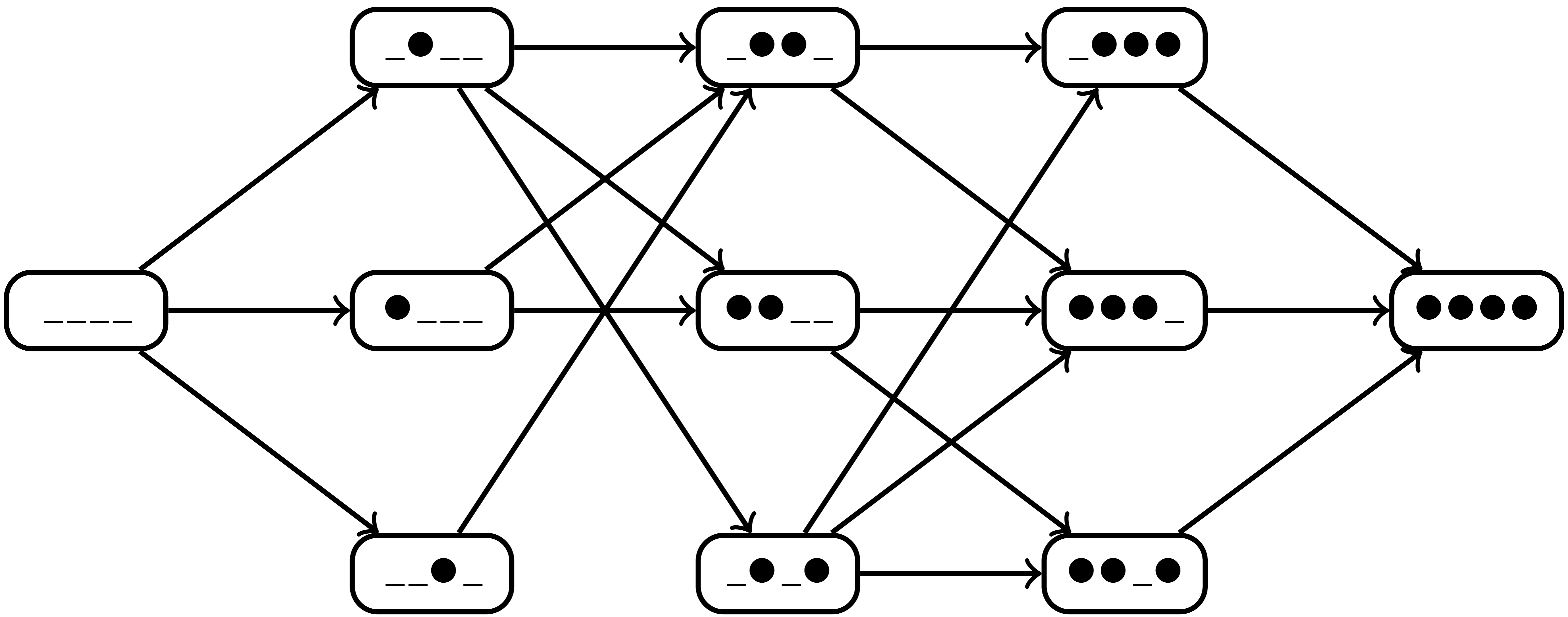
Decoding for phrase-based SMT

言語	は	コミュニケーション	の	道具	で	ある
language		communication	of	a means	is	
language		of communication		is a means		
language is		a means of communication				



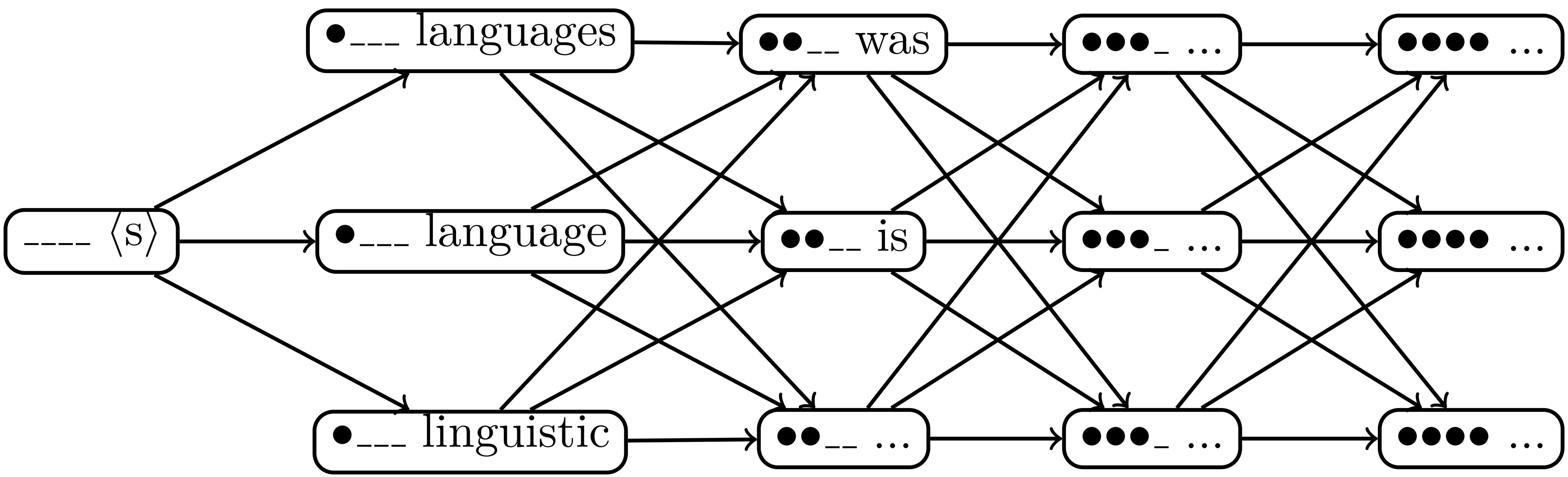
- left-to-right generation + bit-vector for keeping track of covered source positions

Phrase-based decoding



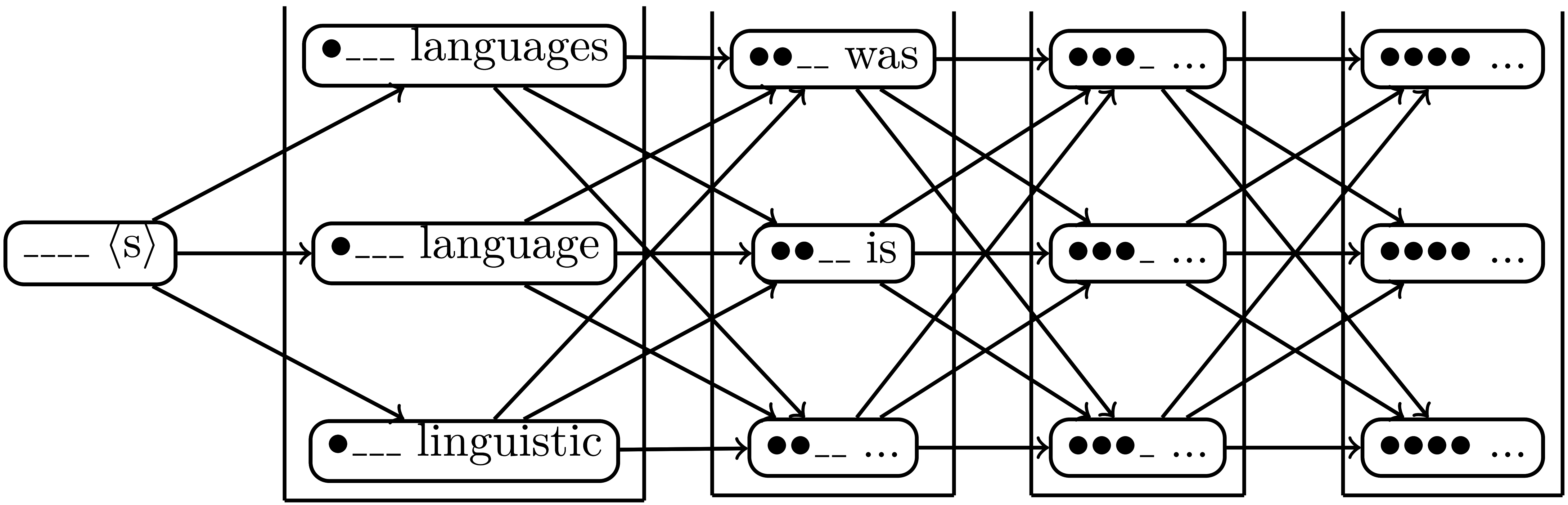
- NP-hard: Traveling salesman problem

Non-local features



- Example: bigram language model
- Enlarged search space

Pruning



- Beam search to limit the search space
  - Multiple stack to keep hypotheses sharing the same # of covered source words

Overview

- Model, Training, Decoding
- Word Alignment
- Phrase-based SMT
- Evaluation
- Optimization

Evaluation

- How do you know translations are good or bad?
- Human judgement
  - Fluency/Adequecy, Human Translation Error Rate (H-TER), Ranking etc.
- Automatic measures: Bleu, Meteor, TER etc.
  - Uses reference translations

Evaluation: ngram precision

Well , I 'd like to stay five nights beginning October twenty-fifth to thirty .

$p_1 = \frac{11}{15}$ $p_2 = \frac{5}{14}$ $p_3 = \frac{3}{13}$ $p_4 = \frac{2}{12}$

- I 'd like to stay there for five nights , from October twenty fifth to the thirtieth .
- I want to stay for five nights , from October twenty fifth to the thirtieth .
- I 'd like to stay for five nights , from October twenty fifth to the thirtieth .
- I would like to reserve a room for five nights , from October twenty fifth to the thirtieth .

Evaluation: BLEU

$$\exp\left(\sum_{n=1}^{N} w_n \log p_n + \min(1 - \frac{r}{c}, 0)\right)$$

- ngram precision: weighted combination
- brevity penalty: penalize too short sentences
  - r = reference length, c = candidate length
- Both factors are computed over the whole document

Overview

- Model, Training, Decoding
- Word Alignment
- Phrase-based SMT
- Evaluation
- Optimization

Optimization: MERT

$$\hat{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}} \sum_{s=1}^{S} l(\operatorname*{argmax}_{\mathbf{e}} \mathbf{w} \cdot \mathbf{h}(\mathbf{e}, \mathbf{f}_s), \mathbf{e}_s)$$

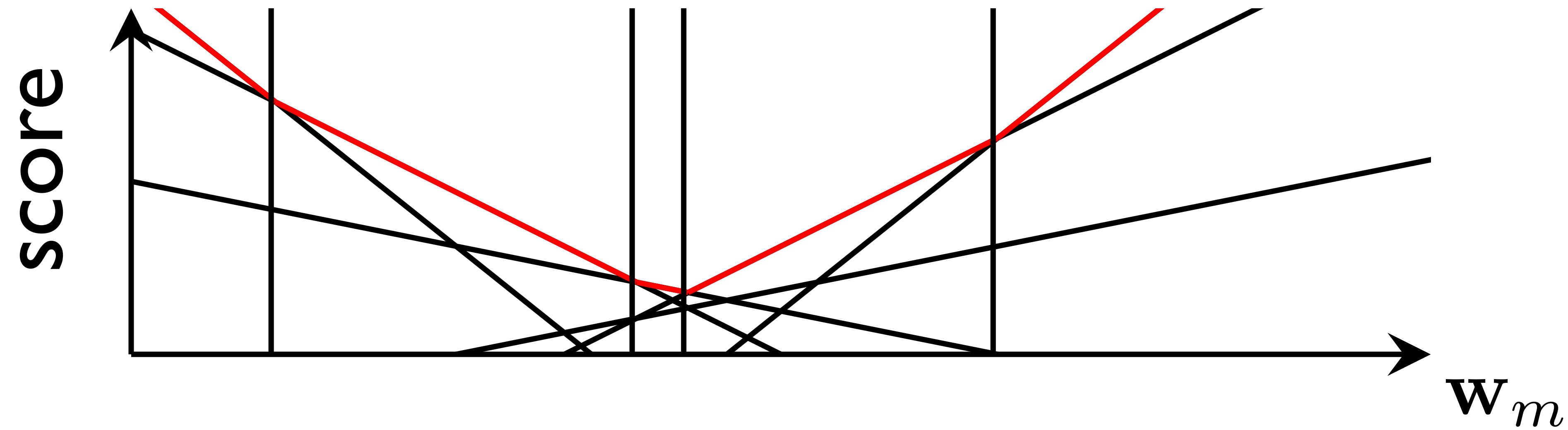
- Minimum Error Rate Training (MERT): directly minimize error (or max-BLEU)
- Small # of real valued features (up to 10?)
- Many local-optima, potential overfitting

MERT

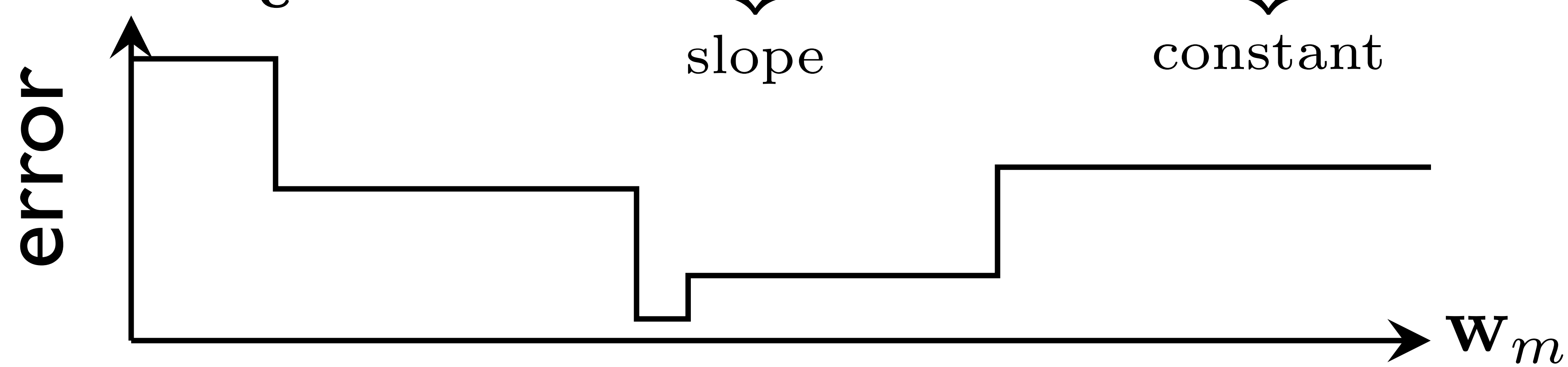
1: **procedure** MERT($\{(\mathbf{e}_s, \mathbf{f}_s)\}_{s=1}^{S}$)
2: **for** $n = 1...N$ **do**
3: Decode and generate nbest list using $\mathbf{w}$
4: Merge nbest list
5: **for** $k = 1...K$ **do**
6: **for** each parameter $m = 1...M$ **do**
7: Solve one dimensional optimization
8: **end for**
9: update $\mathbf{w}$
10: **end for**
11: **end for**
12: **end procedure**

- Generate and merge nbest list across iterations (line 3 and 4)
- Powell's method (or coordinate descent) to perform minimization (line 5-10)

MERT: reduction to 1-dim search



$$\hat{\mathbf{e}} = \operatorname*{argmax}_{\mathbf{e}} \mathbf{w}_m \cdot \underbrace{\mathbf{h}_m(\mathbf{e}, \mathbf{f}_s)}_{\text{slope}} + \underbrace{\mathbf{w}_{m\_} \cdot \mathbf{h}_{m\_}(\mathbf{e}, \mathbf{f}_s)}_{\text{constant}}$$



- If we fix one parameter, it is one dimensional search
- Compute convex hull over a set of lines

MERT: in practice

- Many random starting points (Macherey et al., 2008; Moore and Quirk, 2008)
- Many random directions (Macherey et al., 2008)
- Error count smoothing (Cer et al., 2008)
- Regularization (Hayashi et al., 2009)

Summary

- We quickly reviews basics of SMT:
  - Model, Training, Decoding
  - Word alignment
  - Phrase-based SMT
  - Evaluation
  - Optimization

SMT: Softwares

- GIZA++, gizapp, mgiza: translation model
  - gizapp: http://code.google.com/p/giza-pp/
  - mgiza: http://geek.kyloo.net/software/doku.php
- Alignment by joint training
  - Berkeley Aligner: http://code.google.com/p/berkeleyaligner/
  - PostCAT: http://www.seas.upenn.edu/~strctlrn/CAT/CAT.html
- language models
  - srilm: http://www.speech.sri.com/projects/srilm/
- phrase-based SMT
  - Moses: http://www.statmt.org/moses/

References

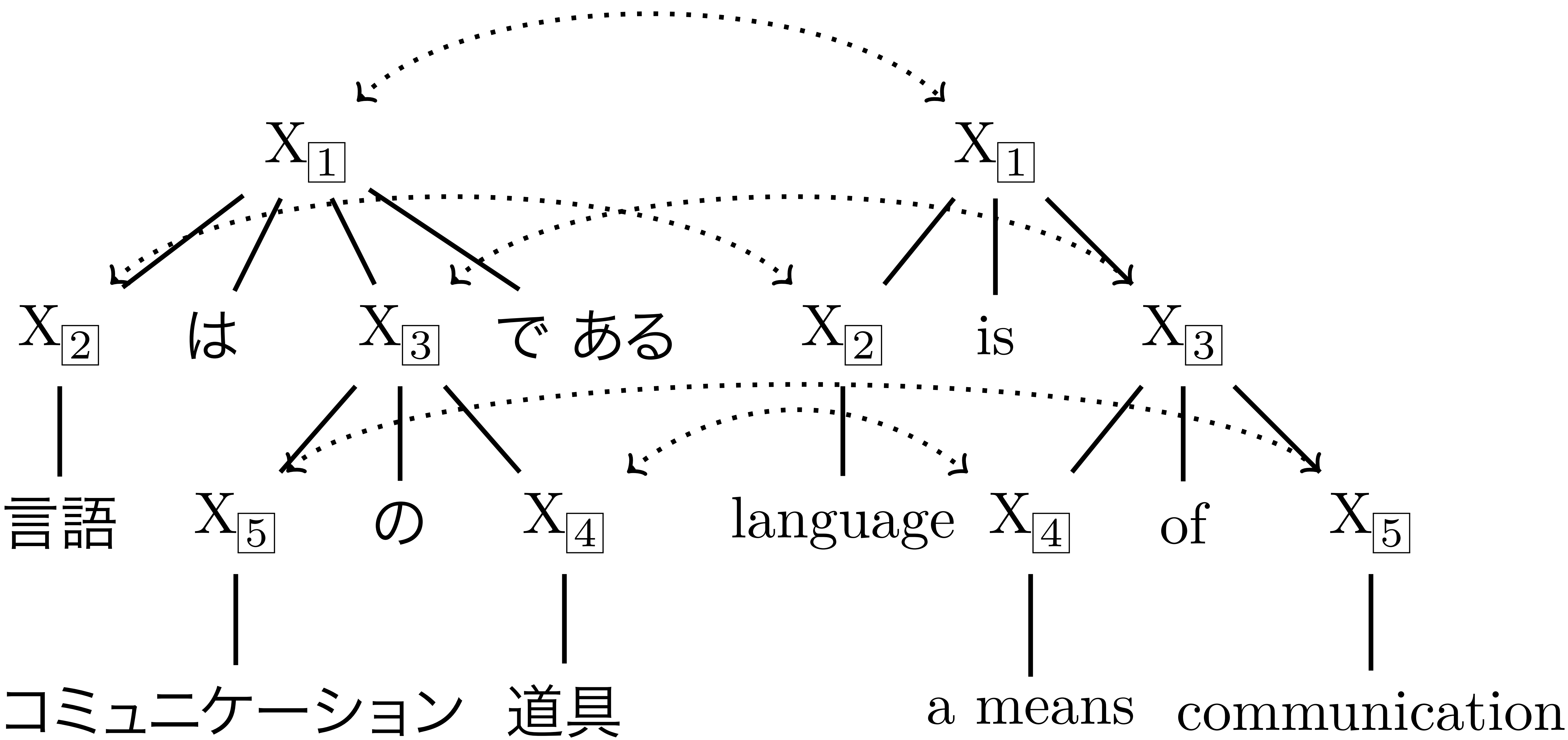
- P. F. Brown, J. Cocke, S. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, ``A statistical approach to machine translation," *Computational Linguistics*, vol. 16, no. 2, pp. 79--85, 1990.
- P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer, ``The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263--311, 1993.
- D. Cer, D. Jurafsky, and C. D. Manning, ``Regularization and search for minimum error rate training," in *Proceedings of the Third Workshop on Statistical Machine Translation*, (Columbus, Ohio), pp. 26--34, Association for Computational Linguistics, June 2008.
- K. Ganchev, J. a. V. Grac ¸a, and B. Taskar, ``Better alignments = better translations?," in *Proceedings of ACL-08: HLT*, (Columbus, Ohio), pp. 986--993, Association for Computational Linguistics, June 2008.
- K. Hayashi, T. Watanabe, H. Tsukada, and H. Isozaki, ``Structural Support Vector Machines for Log-Linear Approach in Statistical Machine Translation," in *Proc. of the International Workshop on Spoken Language Translation*, (Tokyo, Japan), pp. 144--151, 2009.
- K. Knight, ``Decoding complexity in word-replacement translation models," *Comput. Linguist.*, vol. 25, no. 4, pp. 607--615, 1999.

References

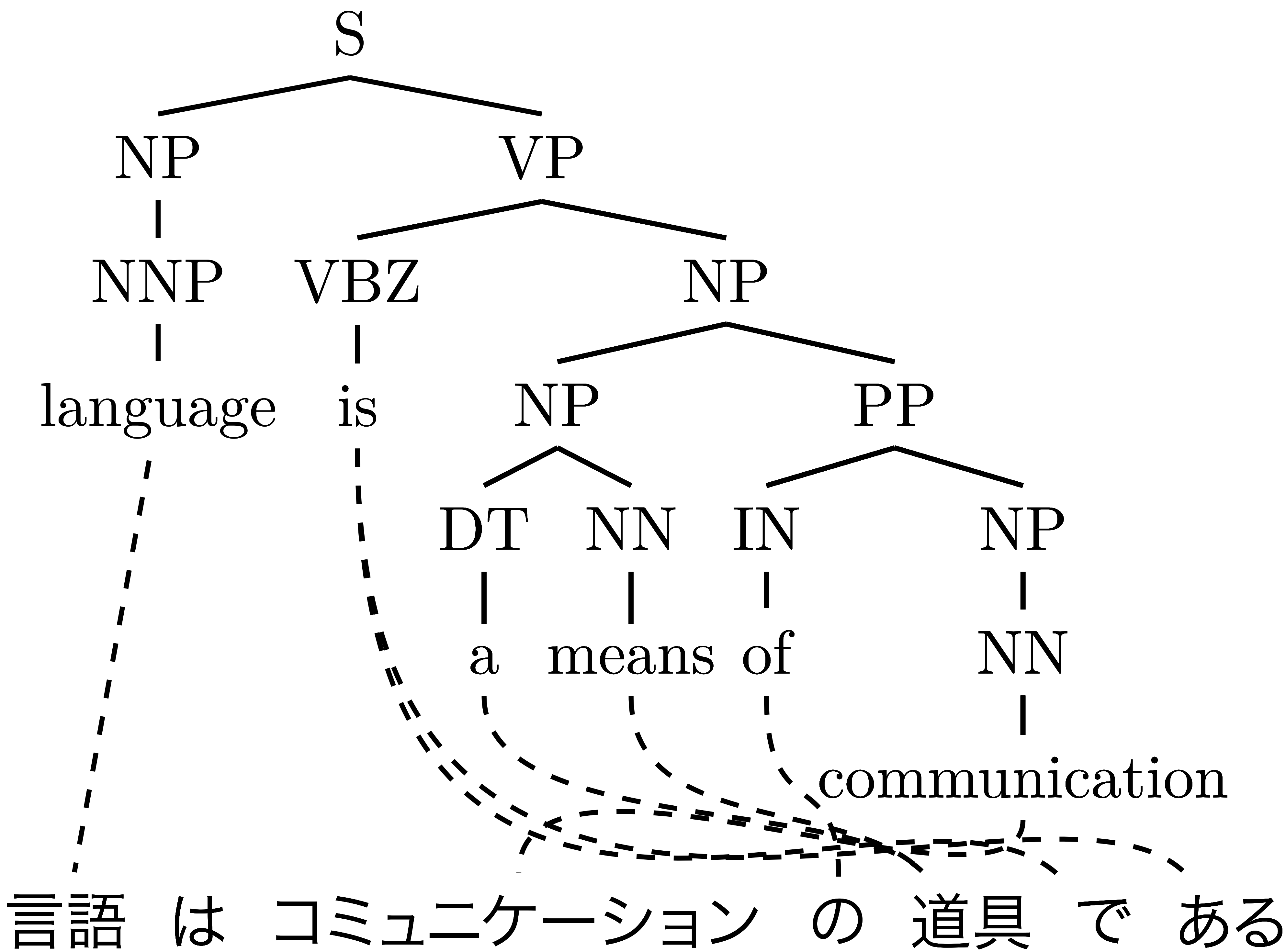
- P. Koehn, F. J. Och, and D. Marcu, ``Statistical pharse-based translation," in *Proc. of HLT-NAACL 2003*, (Edmonton), pp. 48--54, May-June 2003.
- P. Liang, B. Taskar, and D. Klein, ``Alignment by agreement," in *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, (New York City, USA), pp. 104--111, Association for Computational Linguistics, June 2006.
- W. Macherey, F. Och, I. Thayer, and J. Uszkoreit, ``Lattice-based minimum error rate training for statistical machine translation," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, (Honolulu, Hawaii), pp. 725--734, Association for Computational Linguistics, October 2008.
- R. C. Moore and C. Quirk, ``Random restarts in minimum error rate training for statistical machine translation," in *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, (Manchester, UK), pp. 585--592, Coling 2008 Organizing Committee, August 2008.
- F. J. Och and H. Ney, ``The alignment template approach to statistical machine translation," *Comput. Linguist.*, vol. 30, no. 4, pp. 417--449, 2004.
- F. J. Och, ``Minimum error rate training in statistical machine translation," in *Proc. of ACL 2003*, (Sapporo, Japan), pp. 160--167, July 2003.

Tree-baed SMT

Hierarchical Phrase-based SMT



Syntax-based MT



Many variants...

tree	(partial) examples
none	Chiang (2007), Zollman and Venugopal (2006)
source	Huang et al. (2006), Liu et al. (2006), Quirk et al. (2005)
target	Galley et al. (2004), Shen et al. (2008)
both	Ding and Palmer (2005), Liu et al. (2009)

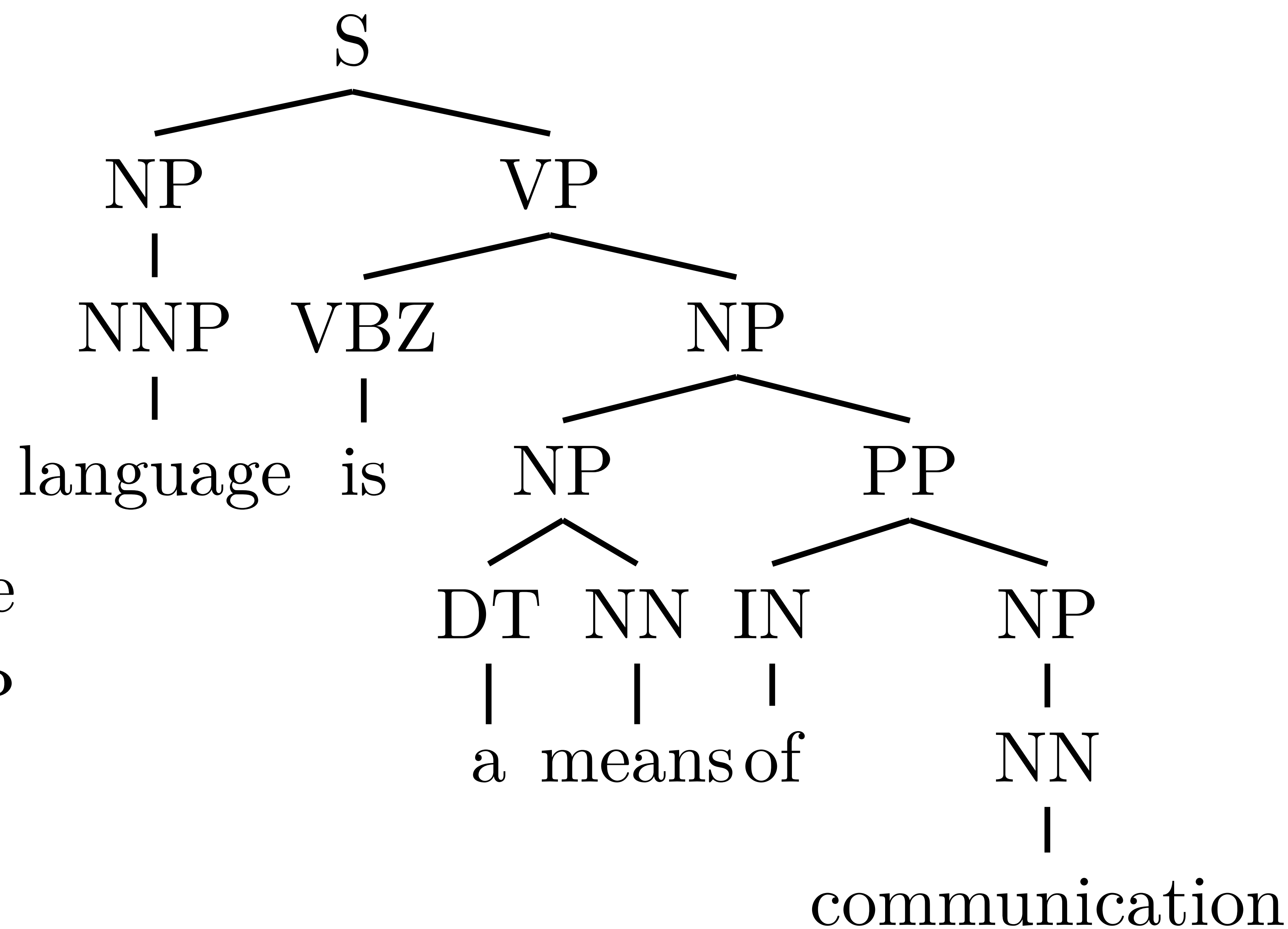
- formally syntactical, linguistically syntactical
- dependency structure and constituency structure
- {tree,string}-to-{tree,string}
- In this talk, we will summarize them as "tree-based MT"

Overview

- Backgrounds
  - CFG, parsing, hypergraph, deductive system, semirings
- Tree-based SMT
  - Synchronous-CFG
  - String-to-Tree/Tree-to-String
  - Bitext parsing

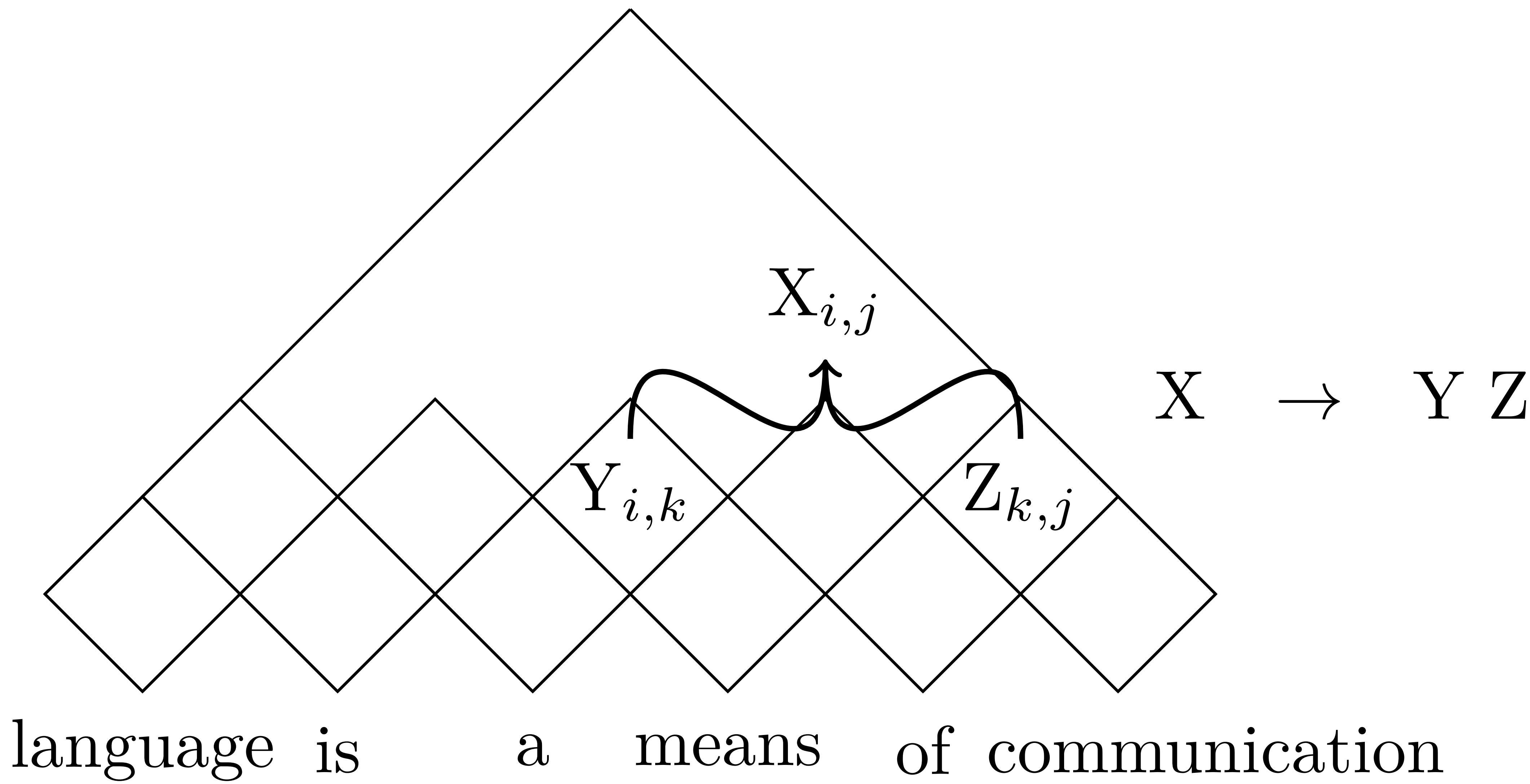
Backgrounds: CFG

$$
\begin{aligned}
\mathrm{S} &\rightarrow \mathrm{NP\ VP} \\
\mathrm{NP} &\rightarrow \mathrm{NNP} \\
\mathrm{NP} &\rightarrow \mathrm{NP\ PP} \\
\mathrm{NP} &\rightarrow \mathrm{DP\ NN} \\
\mathrm{NNP} &\rightarrow \mathrm{language} \\
\mathrm{VP} &\rightarrow \mathrm{VBZ\ NP} \\
\mathrm{VBZ} &\rightarrow \mathrm{is} \\
\mathrm{DT} &\rightarrow \mathrm{a} \\
&\vdots
\end{aligned}
$$



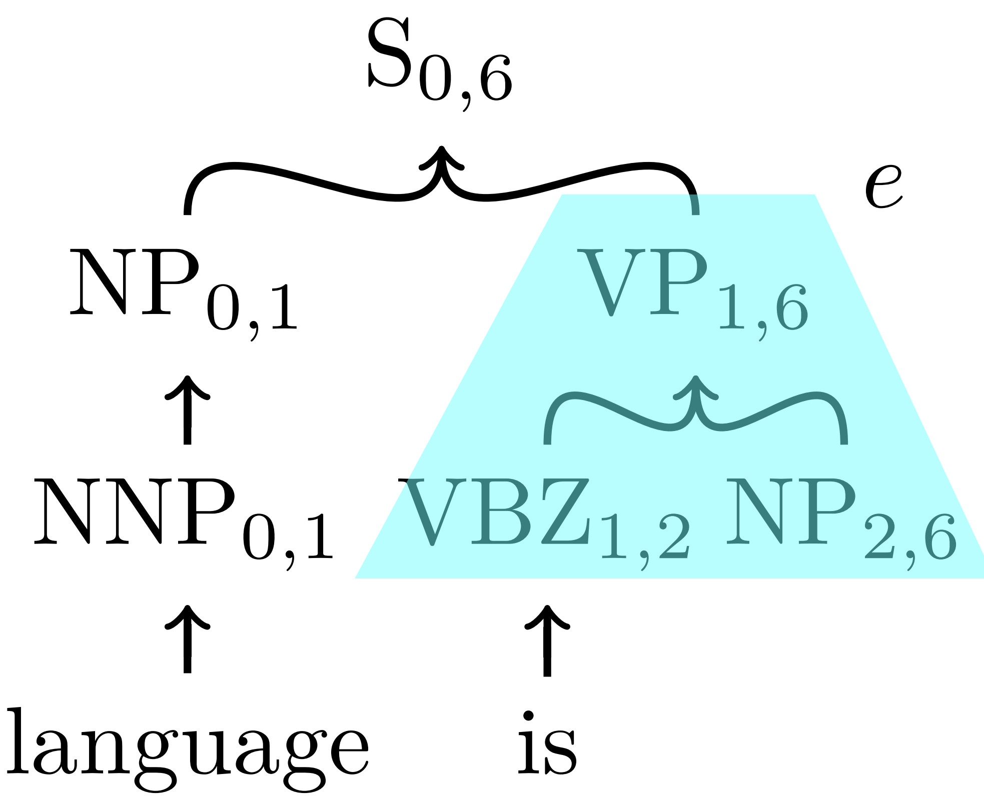
- parsing = intersection problem

Parsing: CKY

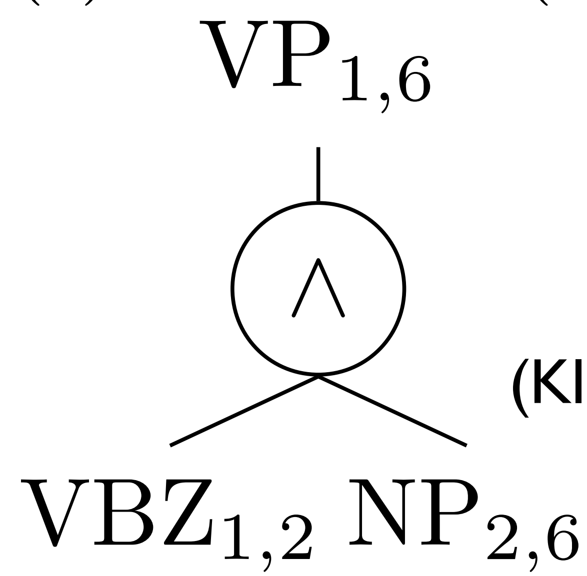


- O(n^3) : For each length n, for each position i, for each rule X → Y Z, for each split point k
- (Bottom-up) topological order

Hypergraph



$$e = \langle \underbrace{\mathrm{VP}_{1,6}}_{h(e)}, \underbrace{\{\mathrm{VBZ}_{1,2}, \mathrm{NP}_{2,6}\}}_{T(e)} \rangle$$



(Klein and Manning, 2001)

- Generalization of graphs:
  - h(e): head node of hyperedge e
  - T(e): tail node(s) of hyperedge e, arity = |T(e)|
  - hyperedge = instantiated rule
- Represented as and-or graphs

Deductive system

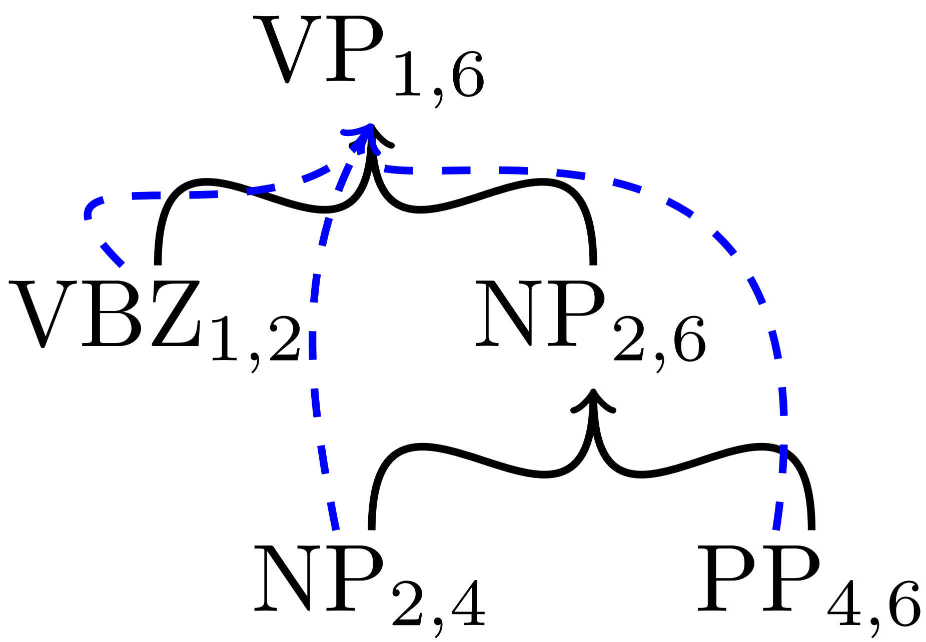
$$\underbrace{\mathrm{VBZ}_{1,2} \quad \mathrm{NP}_{2,6}}_{\downarrow} \qquad \mathrm{VP}_{1,6}$$

$$\frac{\overbrace{\mathrm{VBZ}_{1,2}\ \mathrm{NP}_{2,6}}^{\text{antecedents}}}{\underbrace{\mathrm{VP}_{1,6}}_{\text{consequent}}} \mathrm{VP}_{[i,j]} \rightarrow \mathrm{VBZ}_{[j,k]}\ \mathrm{NP}_{[i,k]}$$

(Shieber et al., 1995)

- Parsing algorithm as a deductive system
- We start from initial items (axioms) until we reach a goal item
- If antecedents are proved, its consequent is proved
- deduction = hyperedge

Packed forest



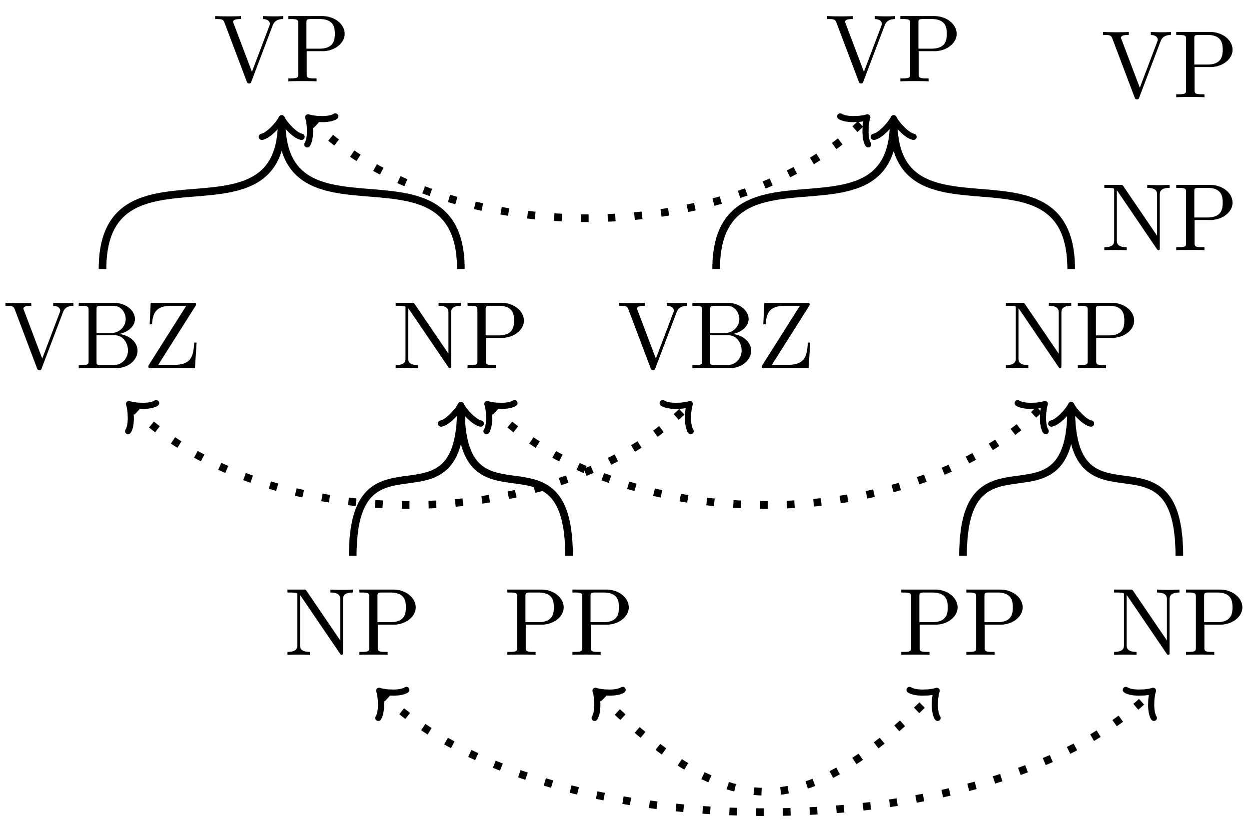
$$\frac{\mathrm{VBZ}_{1,2} \quad \dfrac{\mathrm{NP}_{2,4}\ \ \mathrm{PP}_{4,6}}{\mathrm{NP}_{2,6}}}{\mathrm{VP}_{1,6}}$$

$$\frac{\mathrm{VBZ}_{1,2}\ \ \mathrm{NP}_{2,4}\ \ \mathrm{PP}_{4,6}}{\mathrm{VP}_{1,6}}$$

(Klein and Manning, 2001; Huang and Chiang, 2005)

- A polynomial space encoding of exponentially many parses by sharing common sub-derivations
- Single derivation = tree

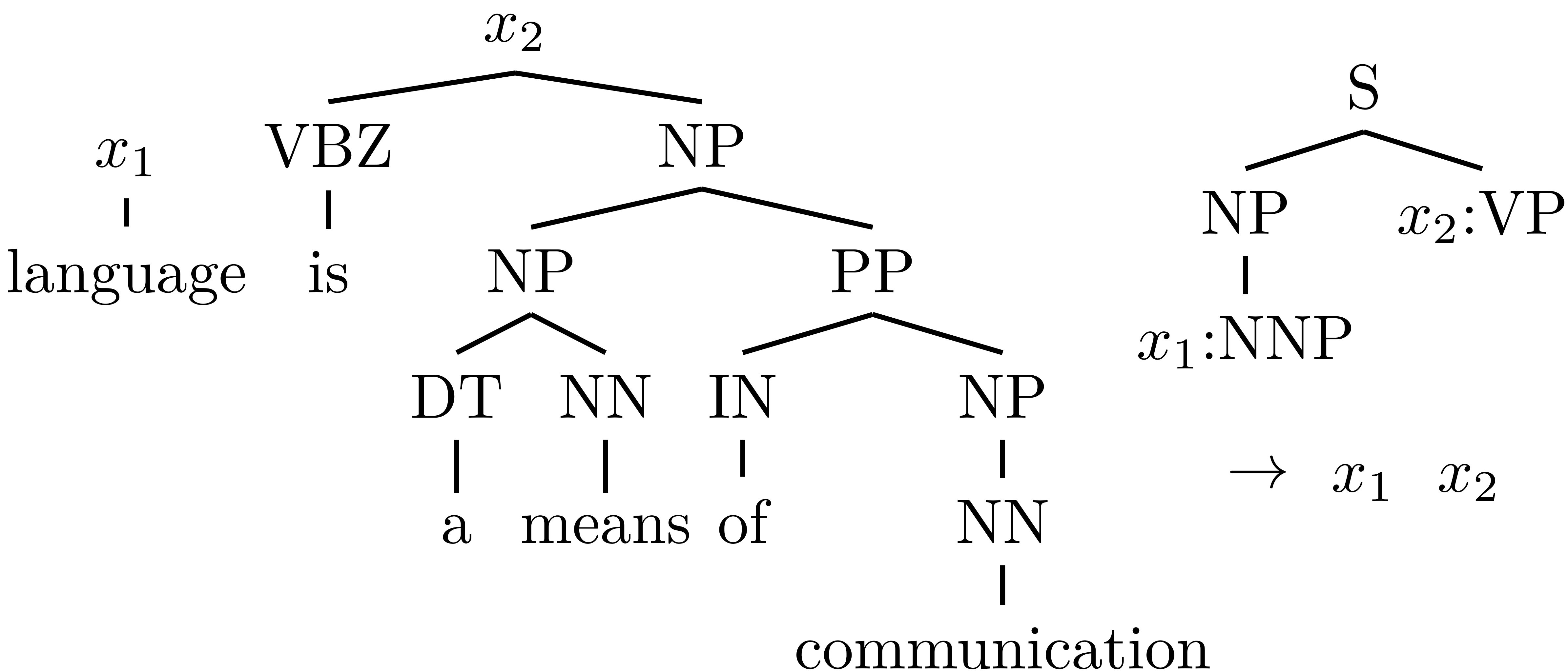
Translation as parsing



$$\begin{aligned} \mathrm{VP} &\rightarrow \langle \mathrm{VBZ}_{\boxed{1}}\ \mathrm{NP}_{\boxed{2}}, \mathrm{VBZ}_{\boxed{1}}\ \mathrm{NP}_{\boxed{2}} \rangle \\ \mathrm{NP} &\rightarrow \langle \mathrm{NP}_{\boxed{1}}\ \mathrm{PP}_{\boxed{2}}, \mathrm{PP}_{\boxed{2}}\ \mathrm{NP}_{\boxed{1}} \rangle \end{aligned}$$

- CFG to synchronous-CFG as in FST with input/output symbols
- Parsing performed over source-yield
- Translation = target-yield of a derivation

Translation as tree-rewrite



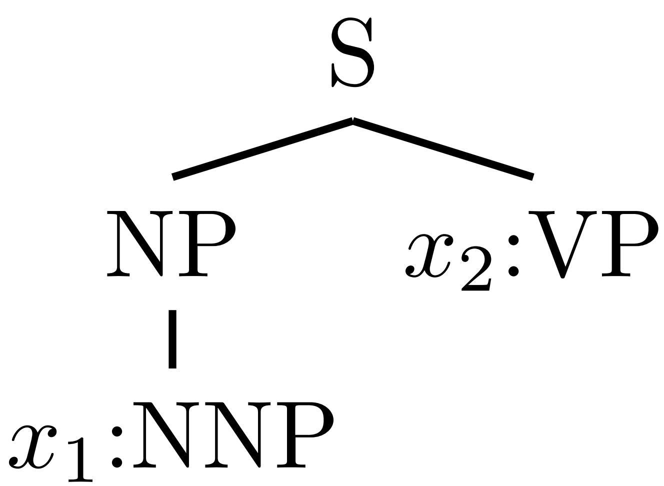
(Galley et al., 2004; Liu et al., 2006; Huang et al., 2006)

- Formalized as tree transducer, tree substitution grammar, or simply, tree-rewrite system
- {tree, string}-to-{tree, string} transformation

Weights and Semirings

$$\mathrm{NP} \xrightarrow{w} \langle \mathrm{NP}_{\boxed{1}}\ \mathrm{PP}_{\boxed{2}}, \mathrm{PP}_{\boxed{2}}\ \mathrm{NP}_{\boxed{1}} \rangle$$

$$\overbrace{\mathrm{NP}_{2,4} : a \quad \mathrm{PP}_{4,6} : b}^{\mathrm{NP}_{2,6} : w \otimes a \otimes b}$$



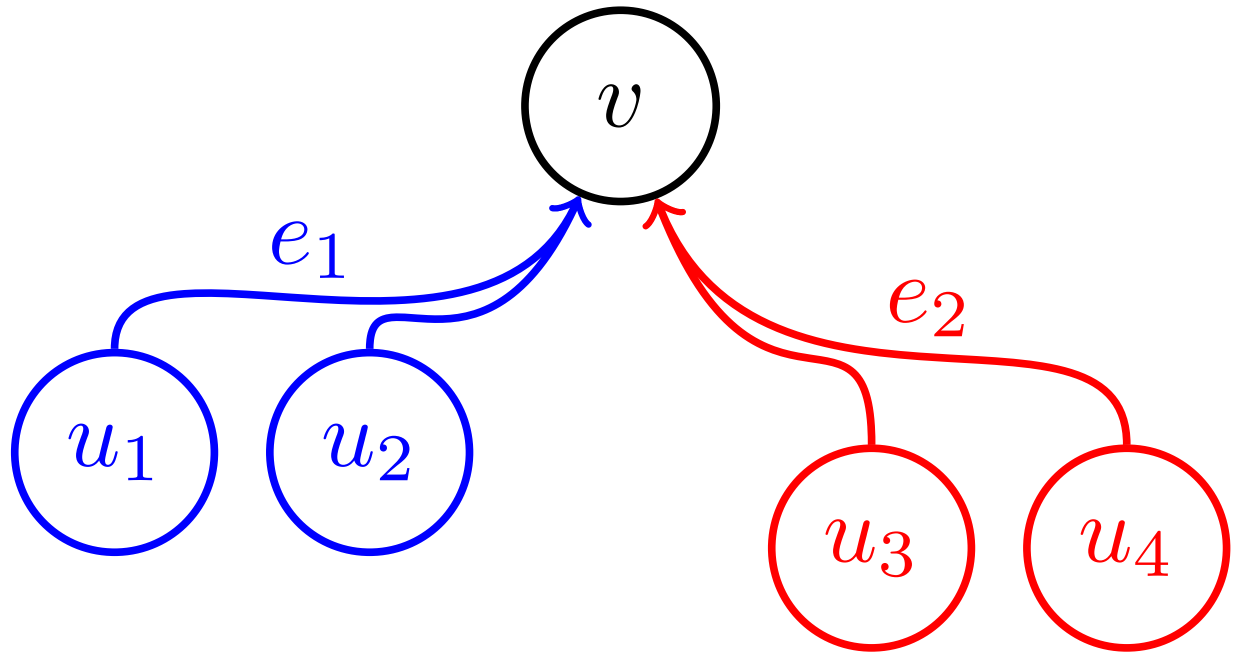
$$\xrightarrow{w} \ x_1 \quad x_2$$

$$\frac{\mathrm{NP}_{2,4} : a \ \ \mathrm{PP}_{4,6} : b}{\mathrm{NP}_{2,6} : w \otimes a \otimes b} : w$$

(Goodman, 1999)

- associate weights as in WFST
- $\otimes$: extension (multiplicative), $\oplus$: summary (additive)

Weights and Semirings



$$
\begin{aligned}
d(v) \quad = \quad & (w(e_1, u_1, u_2) \otimes d(u_1) \otimes d(u_2)) \\
& \oplus (w(e_2, u_3, u_4) \otimes d(u_3) \otimes d(u_4))
\end{aligned}
$$

- The weight of a hyperedge is dependent on antecedents (non-monotonic)
- The weight of a derivation is the product of hyperedge weights
- The weight of a vertex is the summary of (sub-)derivation weights

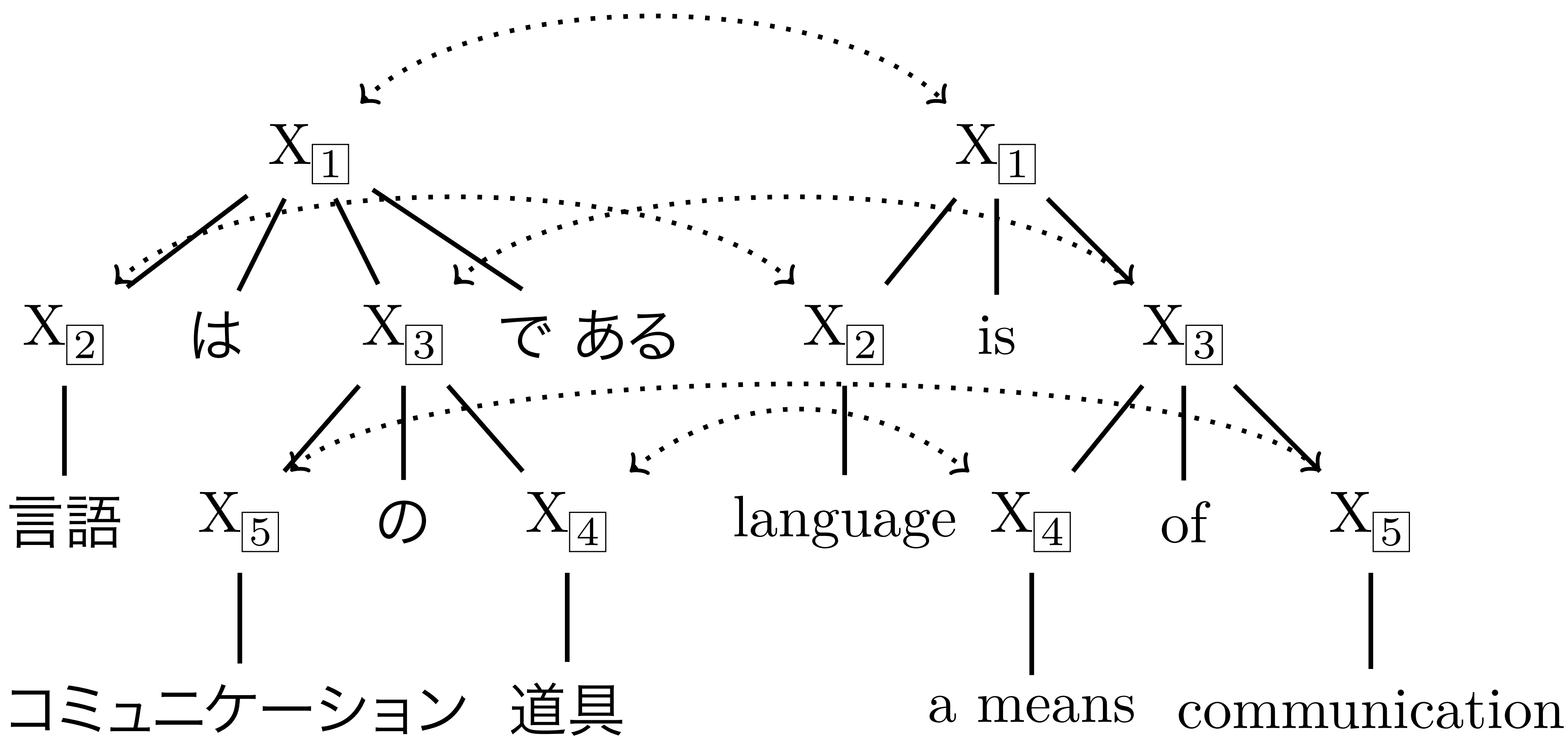
Summary

- Synchronous-CFG: context free rewrite system whose right-hand-side is paired
- Special instances:
  - Inversion Transductive Grammar (ITG) (Wu, 97)
  - Hiero Grammar (Chiang, 2007)
- {tree,string}-to-{tree, string} models
  - Recursive tree rewriting
  - Formalized as tree transducer or tree substitution grammar

Overview

- Backgrounds
  - CFG, parsing, hypergraph, deductive system, semirings
- Tree-based SMT
  - Synchronous-CFG
  - String-to-Tree/Tree-to-String
  - Bitext parsing

Synchronous-CFG



- Derivation: single tree
- Yield: terminals covered by derivation
  - source yield = input sentence
  - target yield = translation

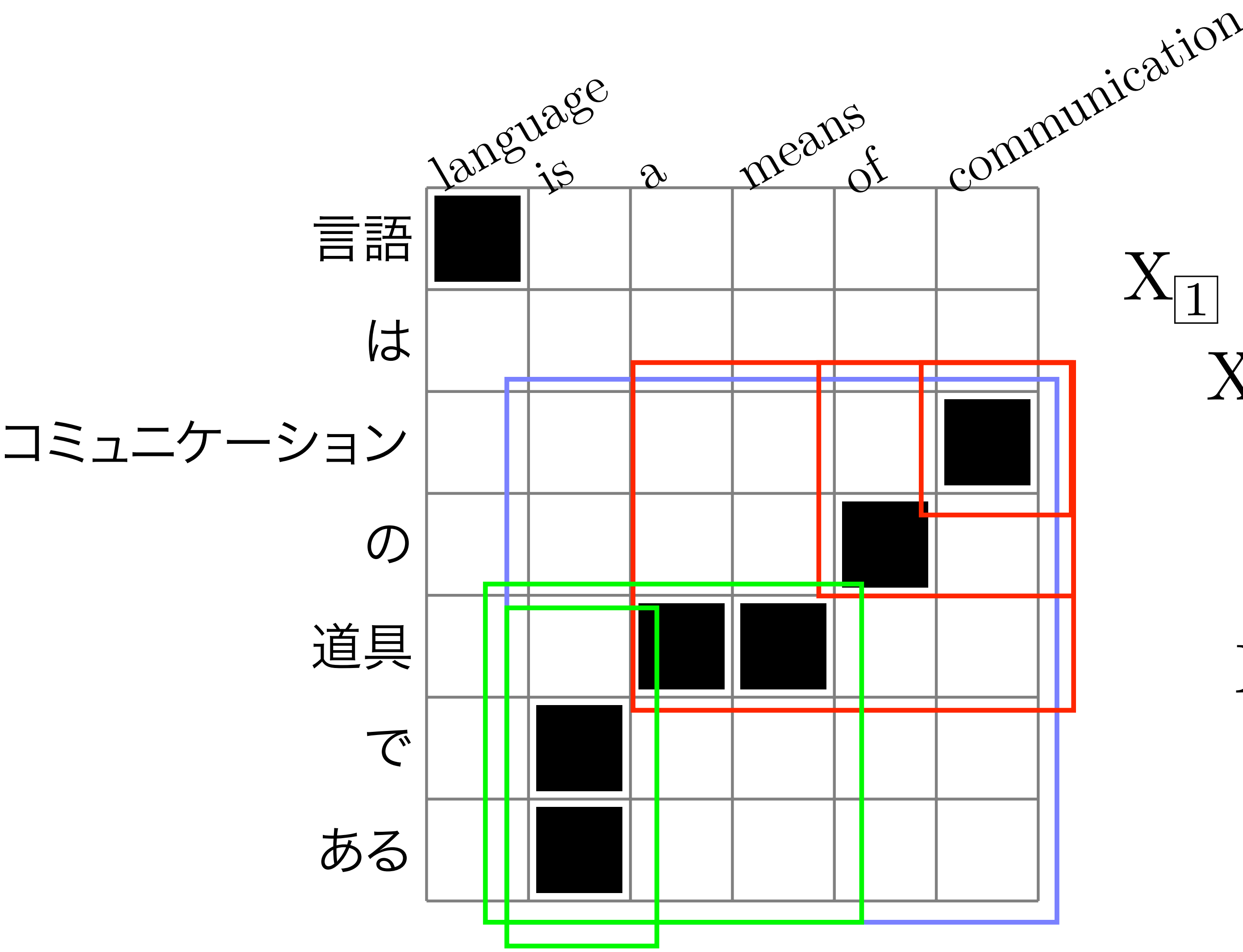
Synchronous-CFG: Model

$$
\begin{aligned}
\mathrm{S} &\rightarrow \langle \mathrm{S}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}}, \mathrm{S}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}} \rangle \\
\mathrm{S} &\rightarrow \langle \mathrm{X}_{\boxed{1}}, \mathrm{X}_{\boxed{1}} \rangle \\
\mathrm{X} &\rightarrow \langle \mathrm{X}_{\boxed{1}}\ \text{の}\ \mathrm{X}_{\boxed{2}}, \mathrm{X}_{\boxed{2}}\ \text{of}\ \mathrm{X}_{\boxed{1}} \rangle \\
\mathrm{X} &\rightarrow \langle \text{道具}, \text{a means} \rangle
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{VP} &\rightarrow \langle \mathrm{VBZ}_{\boxed{1}}\ \mathrm{NP}_{\boxed{2}}, \mathrm{VBZ}_{\boxed{1}}\ \mathrm{NP}_{\boxed{2}} \rangle \\
\mathrm{NP} &\rightarrow \langle \mathrm{NP}_{\boxed{1}}\ \mathrm{PP}_{\boxed{2}}, \mathrm{PP}_{\boxed{2}}\ \mathrm{NP}_{\boxed{1}} \rangle
\end{aligned}
$$

- Use only two categories, S and X (Chiang, 2007)
- Or, borrow linguistic categories from syntactic parse (Zollman and Venugopal, 2006)

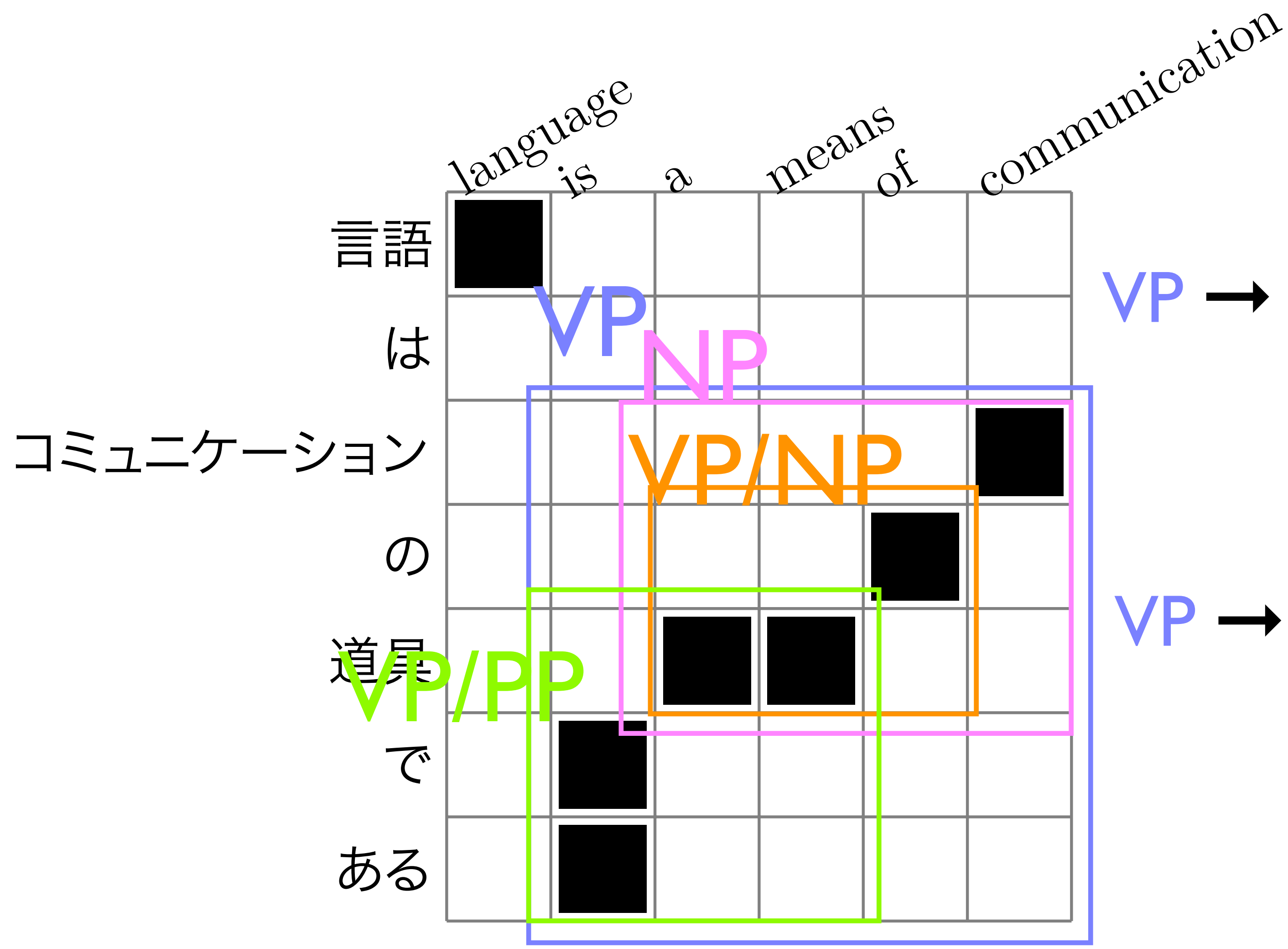
Synchronous-CFG: Extraction



$X_{\boxed{1}}$ の 道具 で ある	is a means of $X_{\boxed{1}}$
$X_{\boxed{1}}$ 道具 で ある	is a means $X_{\boxed{1}}$
$X_{\boxed{1}}$ で ある	is $X_{\boxed{1}}$
$X_{\boxed{1}}$ の $X_{\boxed{2}}$	$X_{\boxed{2}}$ of $X_{\boxed{1}}$
$X_{\boxed{1}}$ の 道具 $X_{\boxed{2}}$	$X_{\boxed{2}}$ a means of $X_{\boxed{1}}$

- From word alignment annotated data, extract phrases
- Sub-phrases treated as non-terminal

Synchronous-CFG: Extraction



VP → コミュニケーション VP/NP で ある

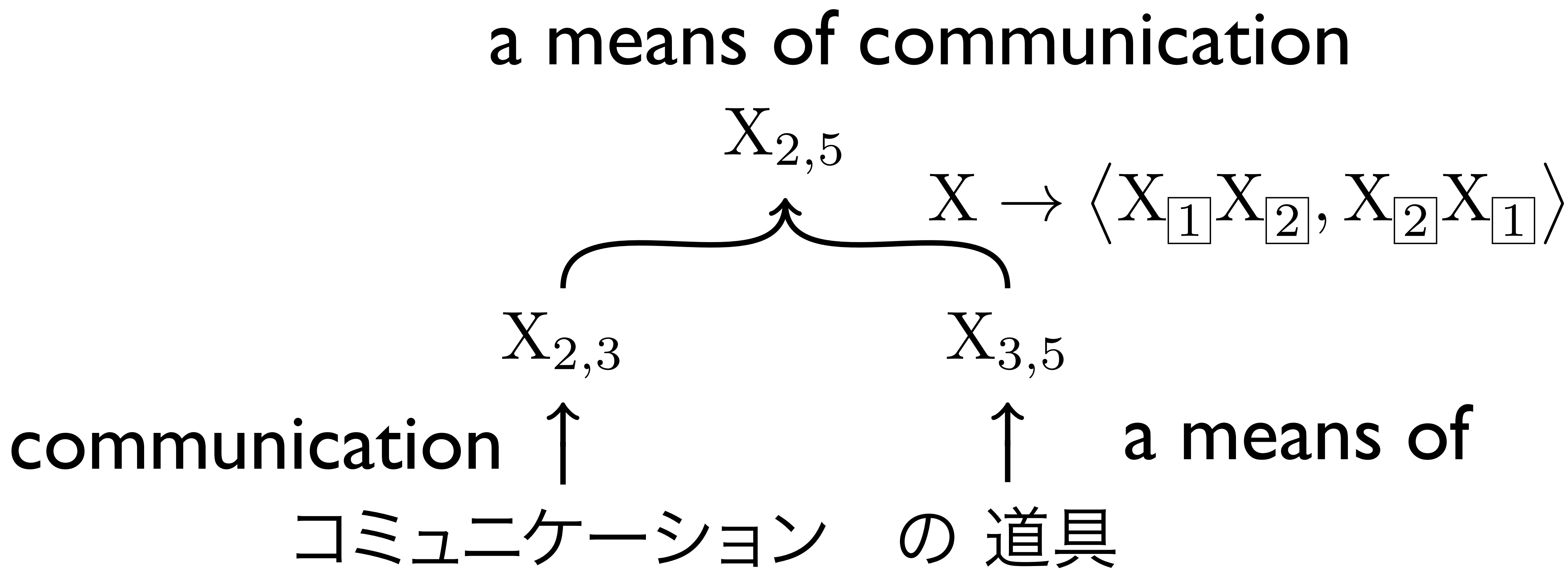
is VP/NP communication

VP → コミュニケーション の VP/PP

VP/PP of communication

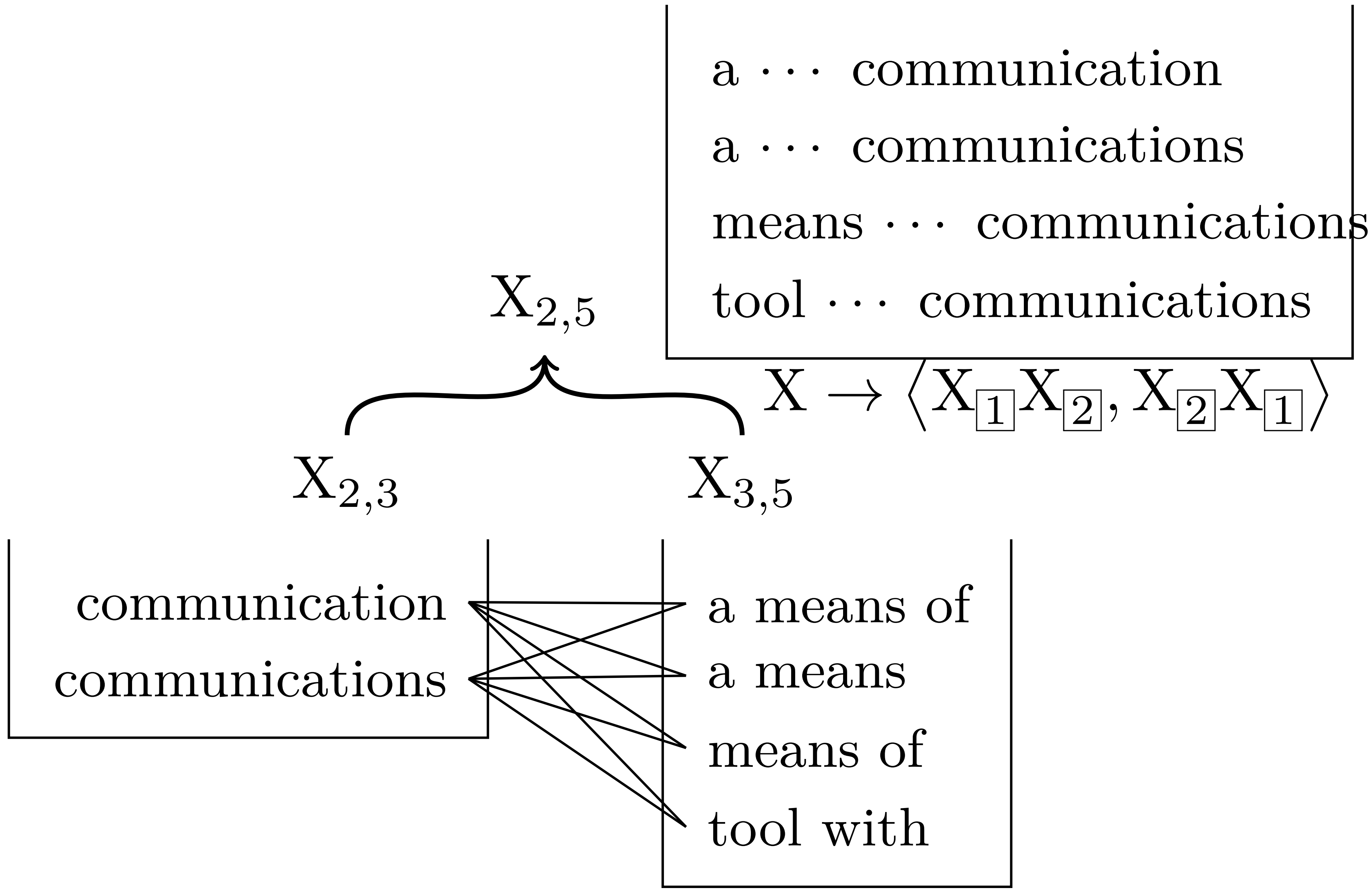
- Borrow syntactic categories eitehr from souce or target parse tree
- When no syntactil categories assigned:
  - Try combination(+) or subtraction(/ or \) as in Combinational Category Grammar (CCG)

Synchronous-CFG: Parsing



- translation with SCFG = monolingual parsing
- Parse the input with the source side, build projected target side in parallel
- Complexity: the same as CKY algorithm: O(n^3)

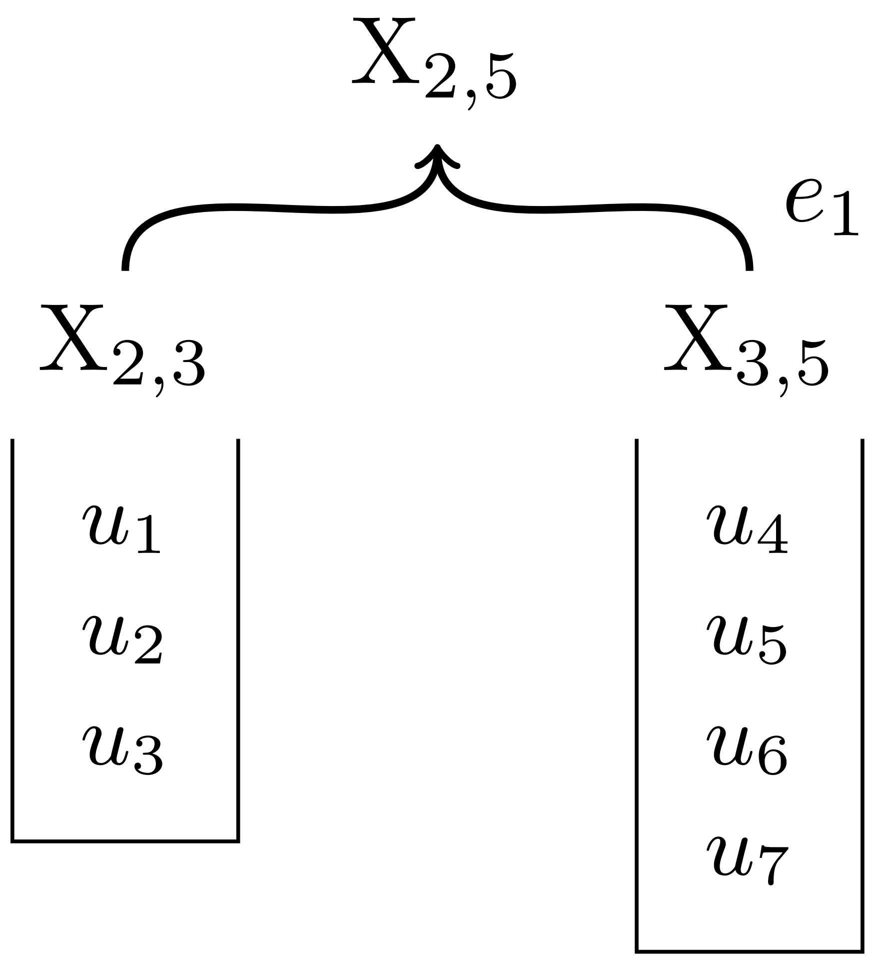
Parsing with non-local features

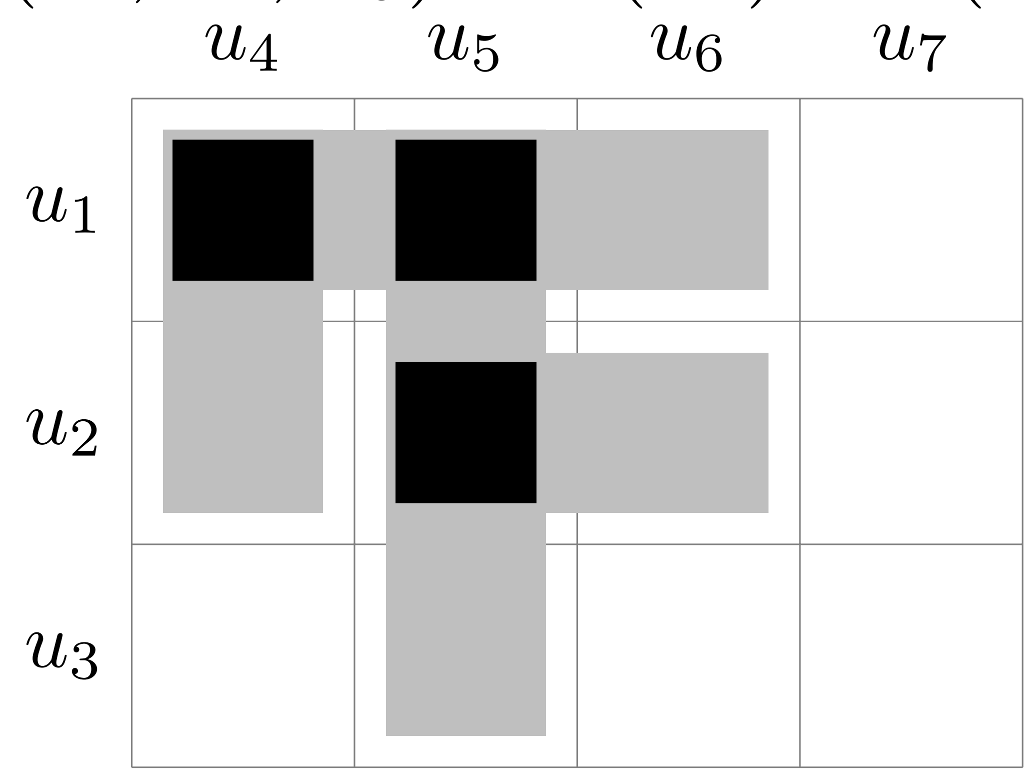


- As in phrase decoding with non-local features (i.e. ngram), it is the same as the CKY algorithm with enlarged search space

Cube Pruning: Basics

$$w(e_1, u_1, u_4) \otimes d(u_1) \otimes d(u_4)$$
$$w(e_1, u_1, u_5) \otimes d(u_1) \otimes d(u_5)$$
$$w(e_1, u_2, u_5) \otimes d(u_2) \otimes d(u_5)$$

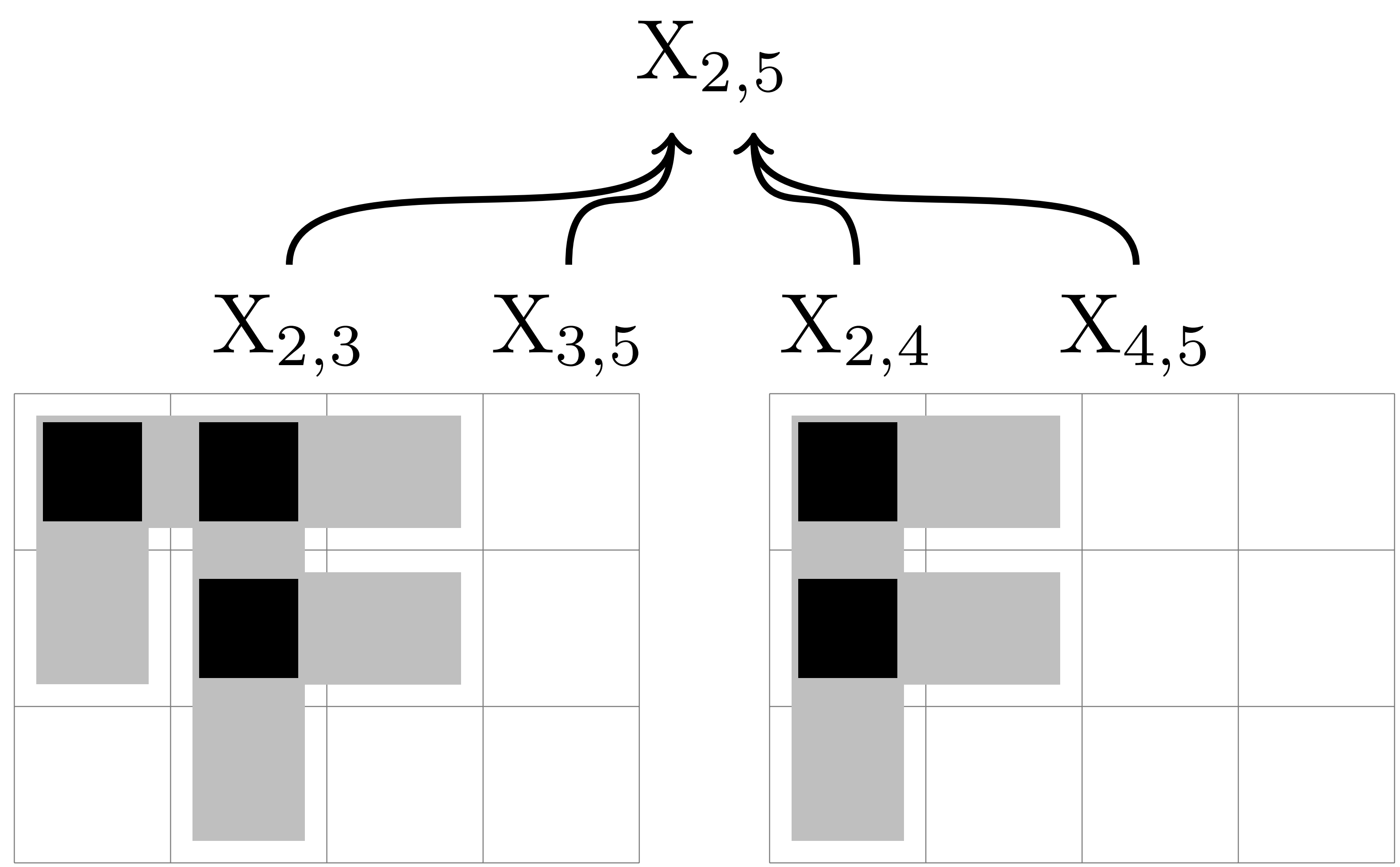




(Chiang, 2007; Huang and Chiang, 2007)

- Lazily enumerate top most items
  - vertices are sorted according to its score
  - pop an item from a priority queu, then expand

Cube Pruning: Grouping



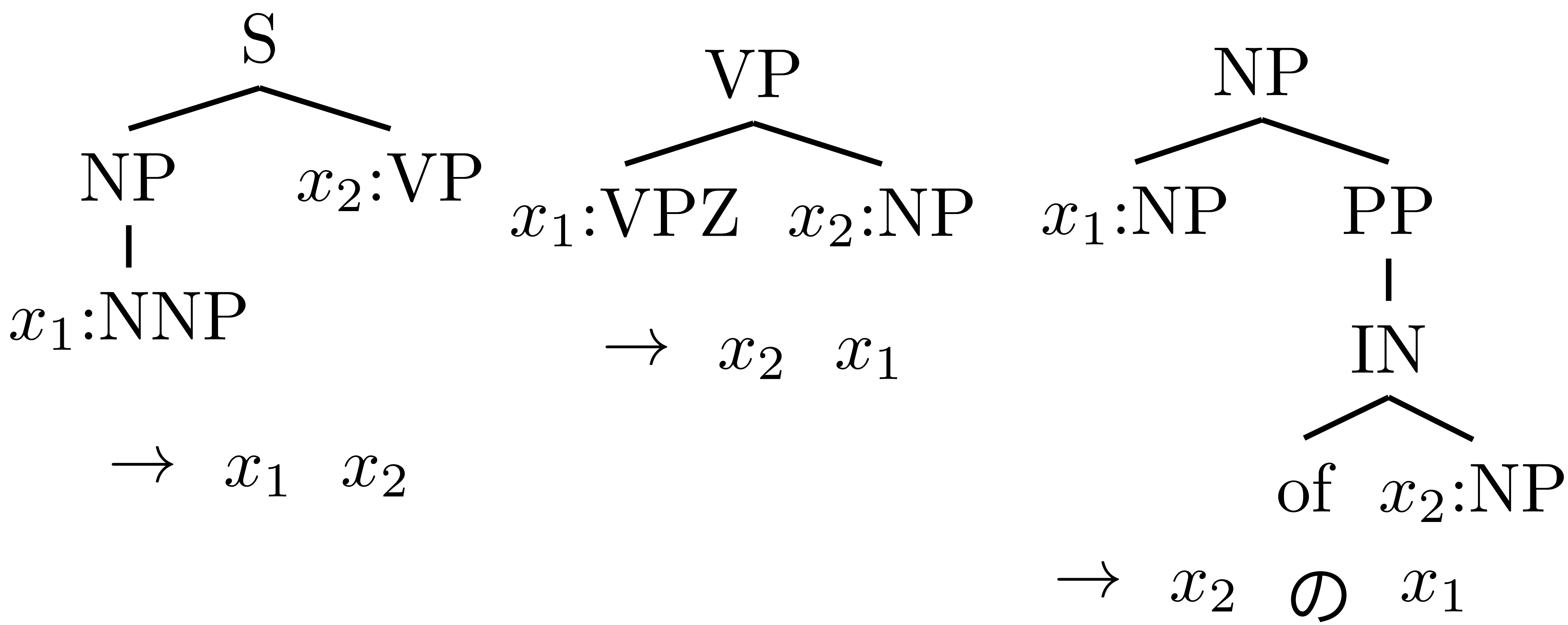
(Chiang, 2007; Huang and Chiang, 2007)

- Simultaneously process the rules sharing the same rhs and span by placing "cubes" in a priority queue

Overview

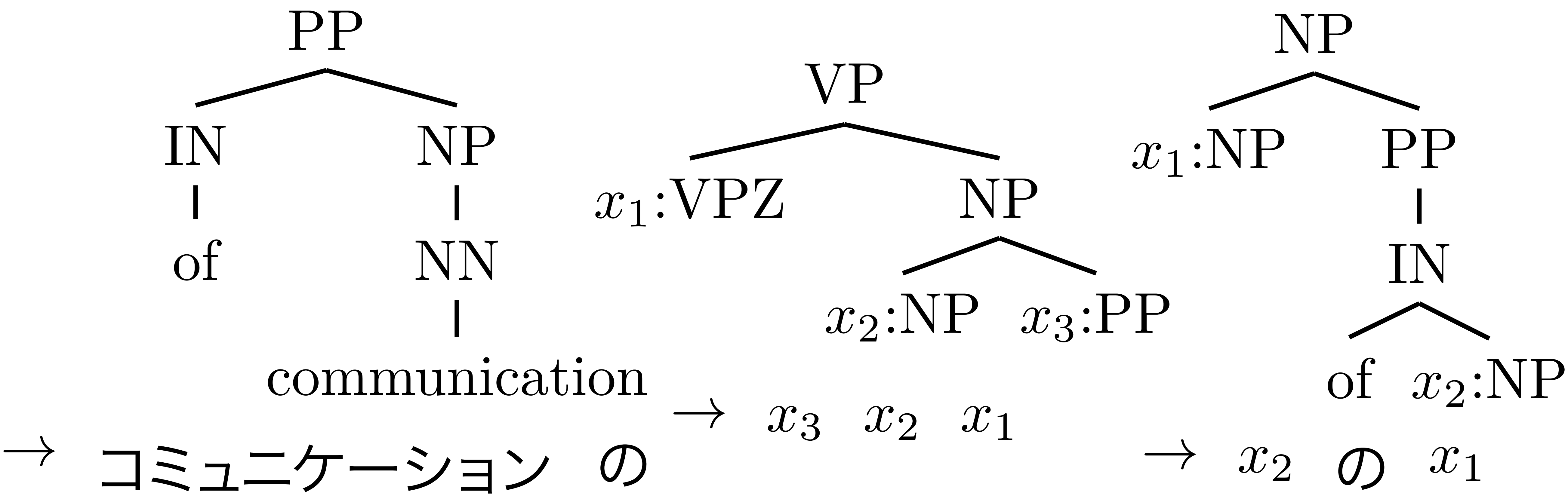
- Backgrounds
  - CFG, parsing, hypergraph, deductive system, semirings
- Tree-based SMT
  - Synchronous-CFG
  - String-to-Tree/Tree-to-String
  - Bitext parsing

{Tree, String}-to-{Tree, String}



- Tree rewriting rules: each rule consists of (sub-)tree structures
- Flat structure = synchronous-CFG

Rules

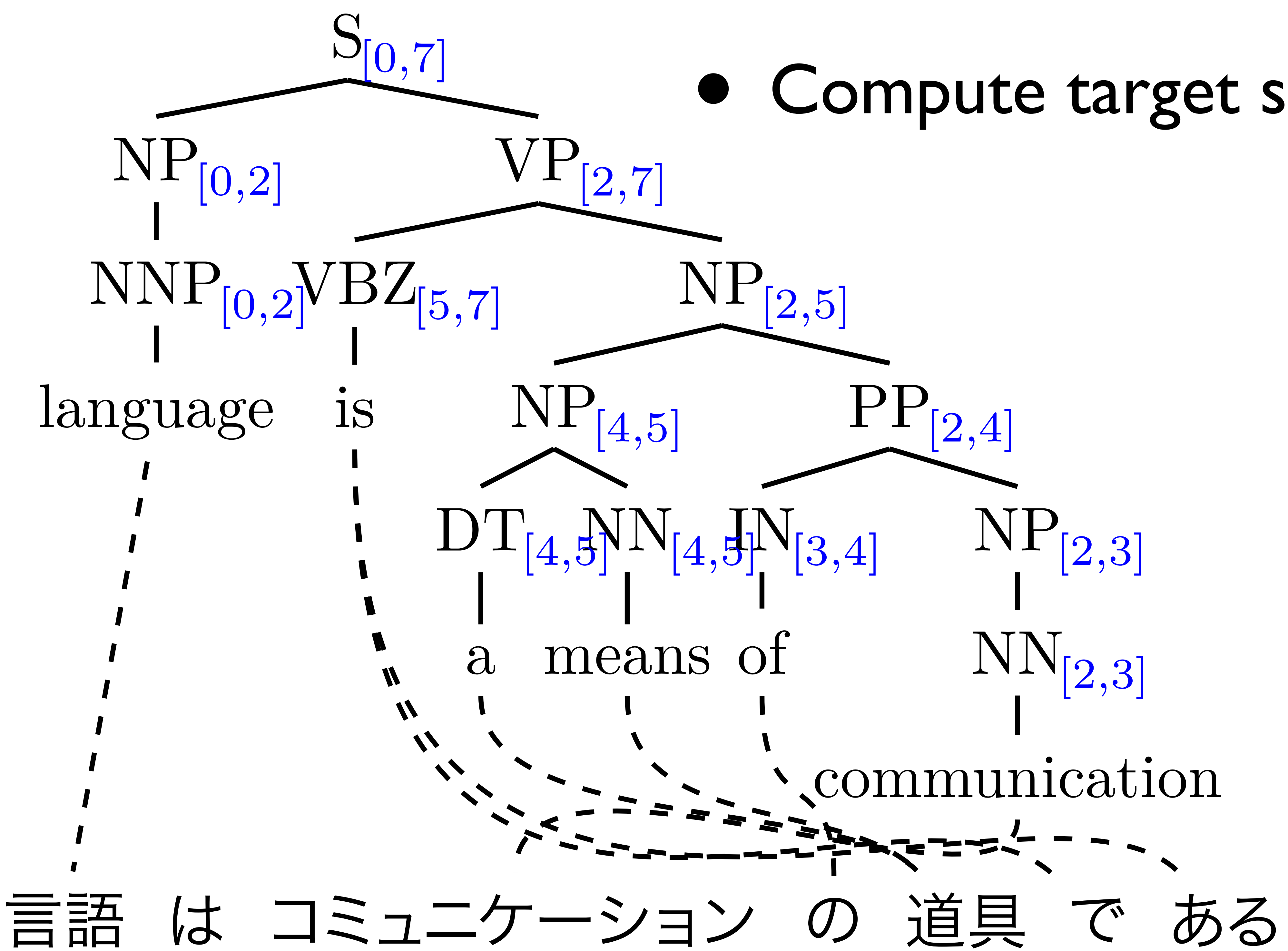


(Galley et al., 2004)

- We can handle various transfer rules:
  - phrasal translation, non-constituent phrase, non-contiguous phrase, insertion/deletion, multi-level reordering, lexicalized reordering, long distance reordering, etc.

Rule extraction

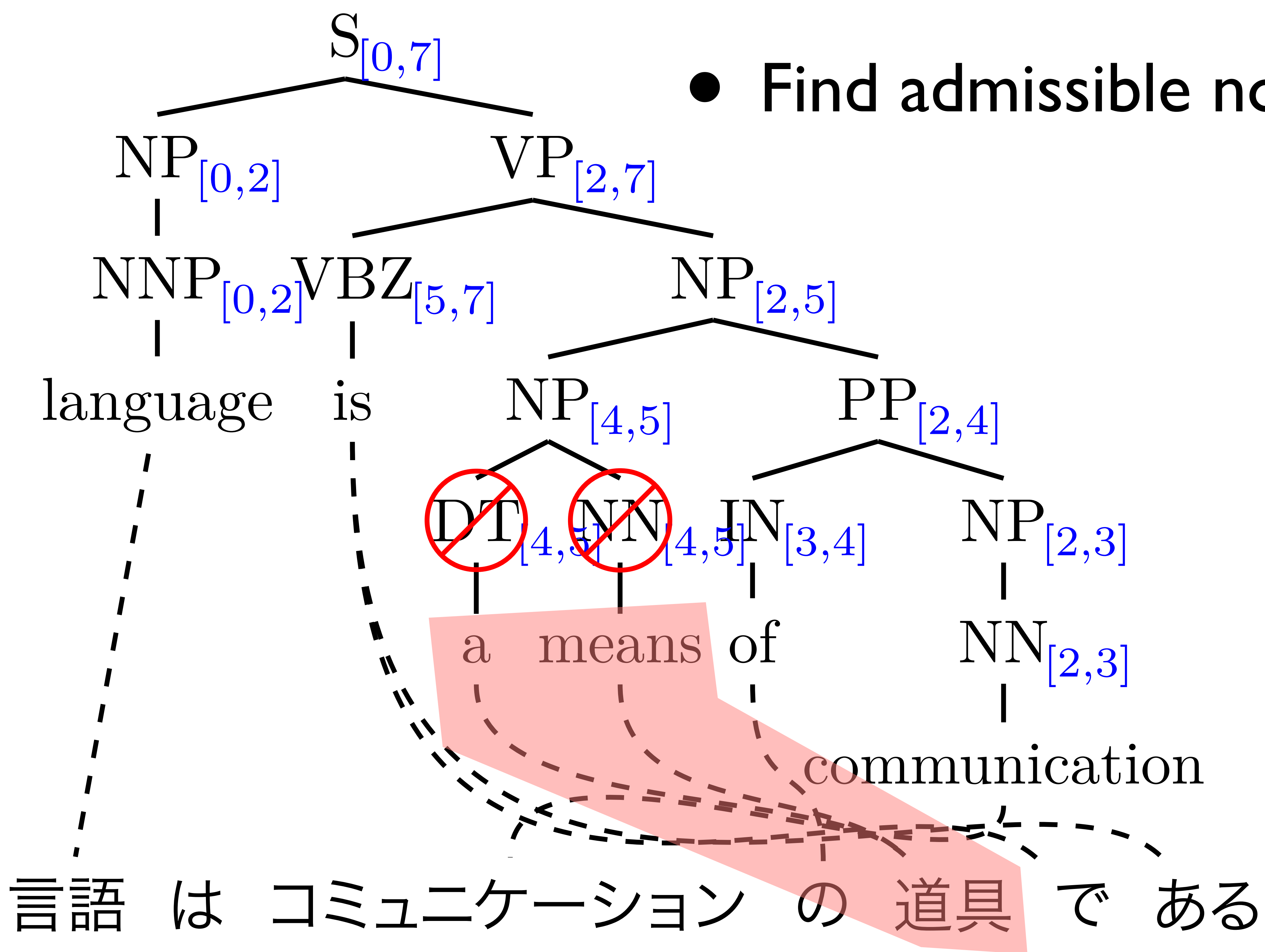
- Compute target spans



(Galley et al., 2004)

Rule extraction

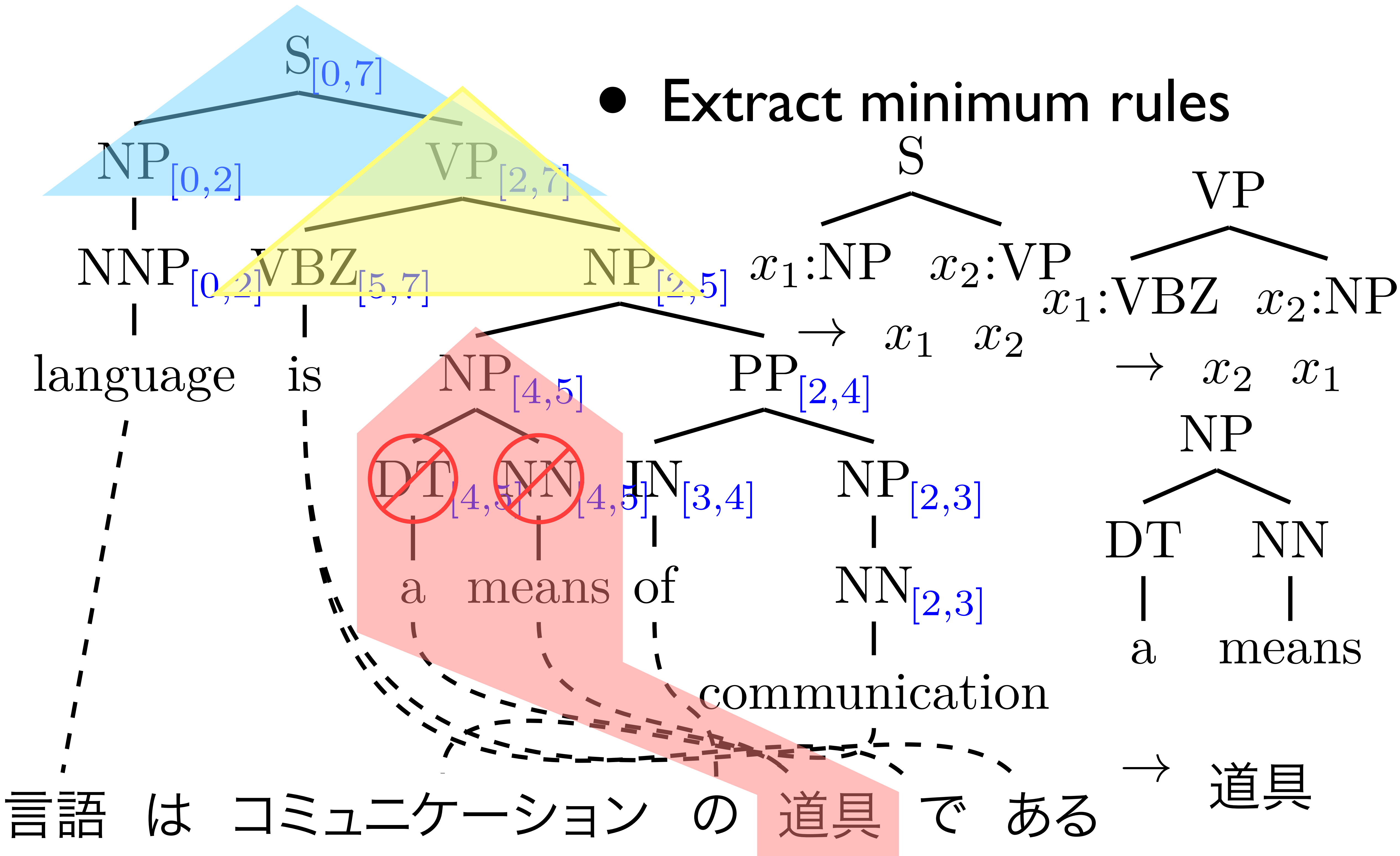
- Find admissible nodes



(Galley et al., 2004)

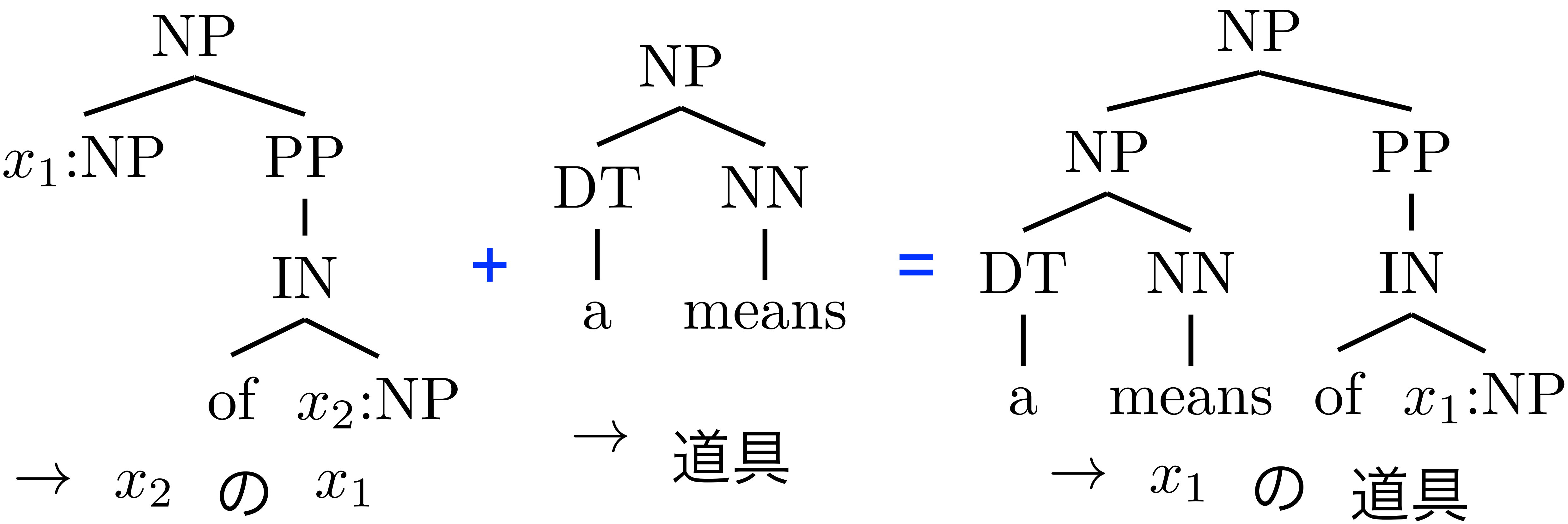
Rule extraction

- Extract minimum rules



(Galley et al., 2004)

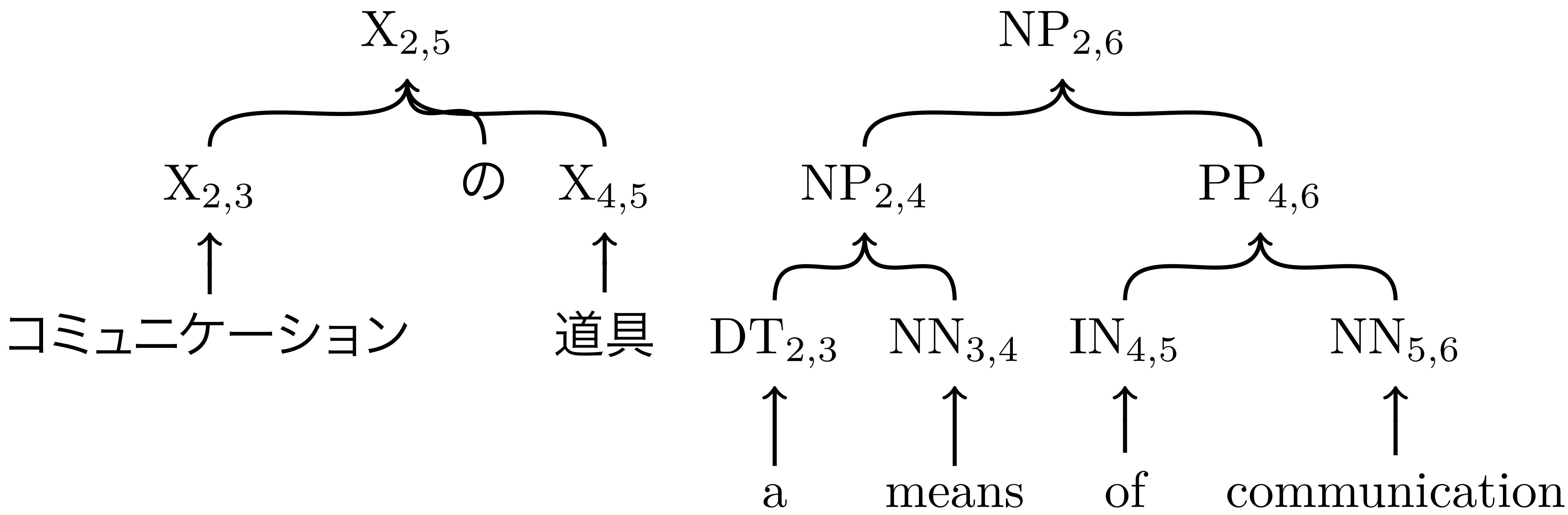
Compound rules



- Tree substitution for compound rules, like phrases from a sequence of words

(Galley et al., 2006)

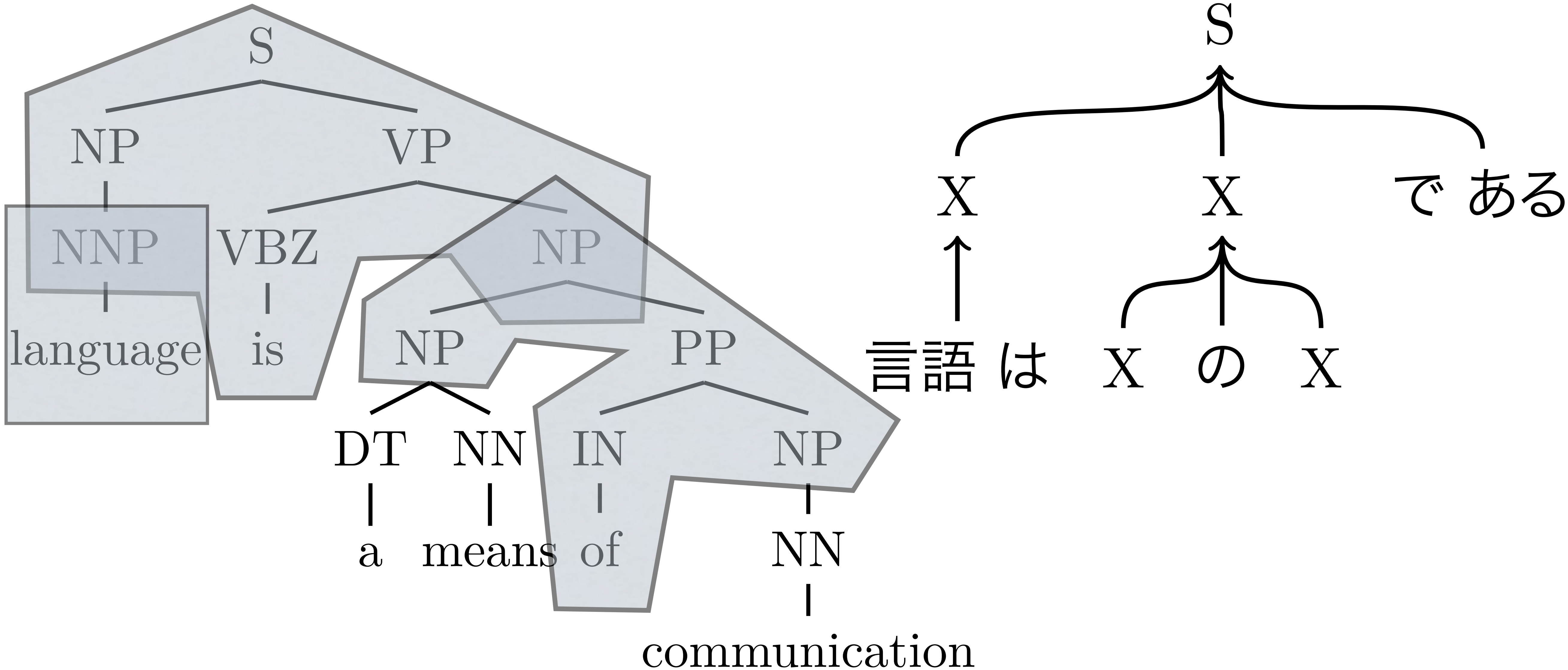
String-to-{string, tree} decoding



(Galley et al., 2004; Huang and Chiang, 2007)

- Similar to SCFG: use flipped string side to perform CKY parsing
- After parsing, tree-reranking from forest

Tree-to-{string, tree} decoding



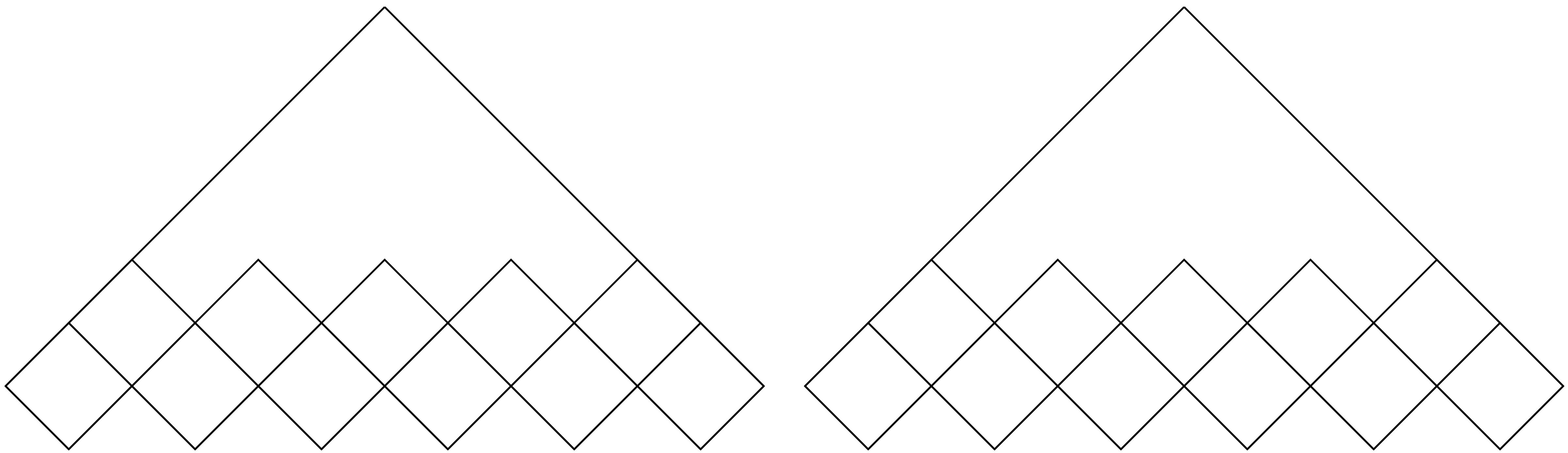
(Huang et al., 2006; Liu et al., 2006)

- Recursively transform by pattern matching over tree
- After matching, forest is rescored (Huang and Chiang; 2007)

Overview

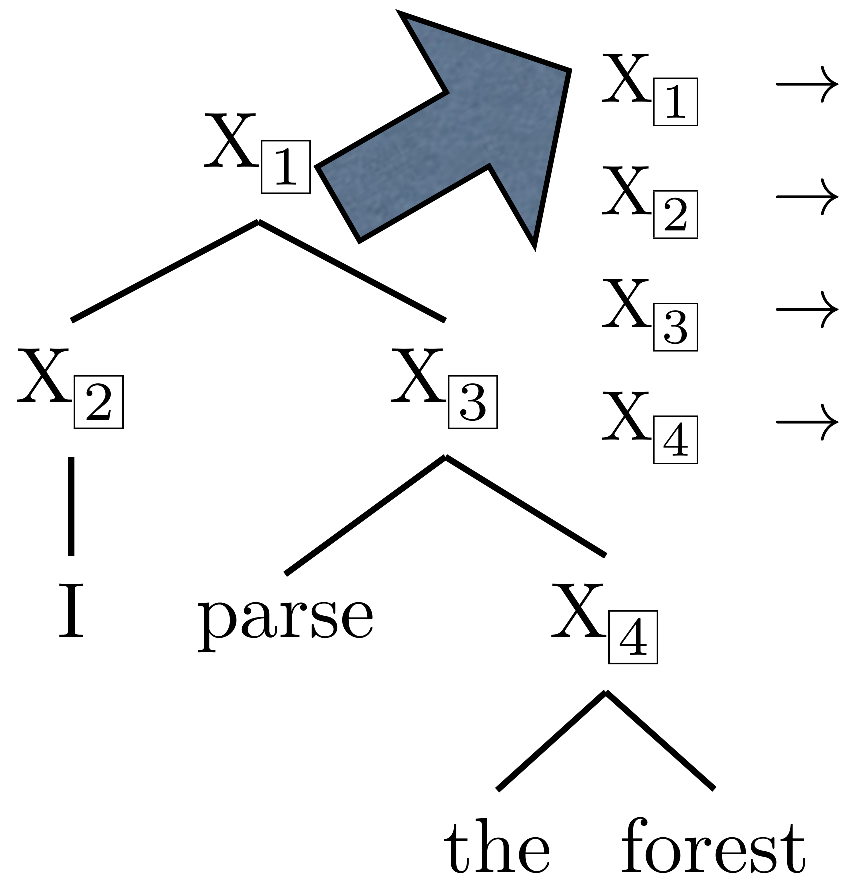
- Backgrounds
  - CFG, parsing, hypergraph, deductive system, semirings
- Tree-based SMT
  - Synchronous-CFG
  - String-to-Tree/Tree-to-String
  - Bitext parsing

Bitext parsing

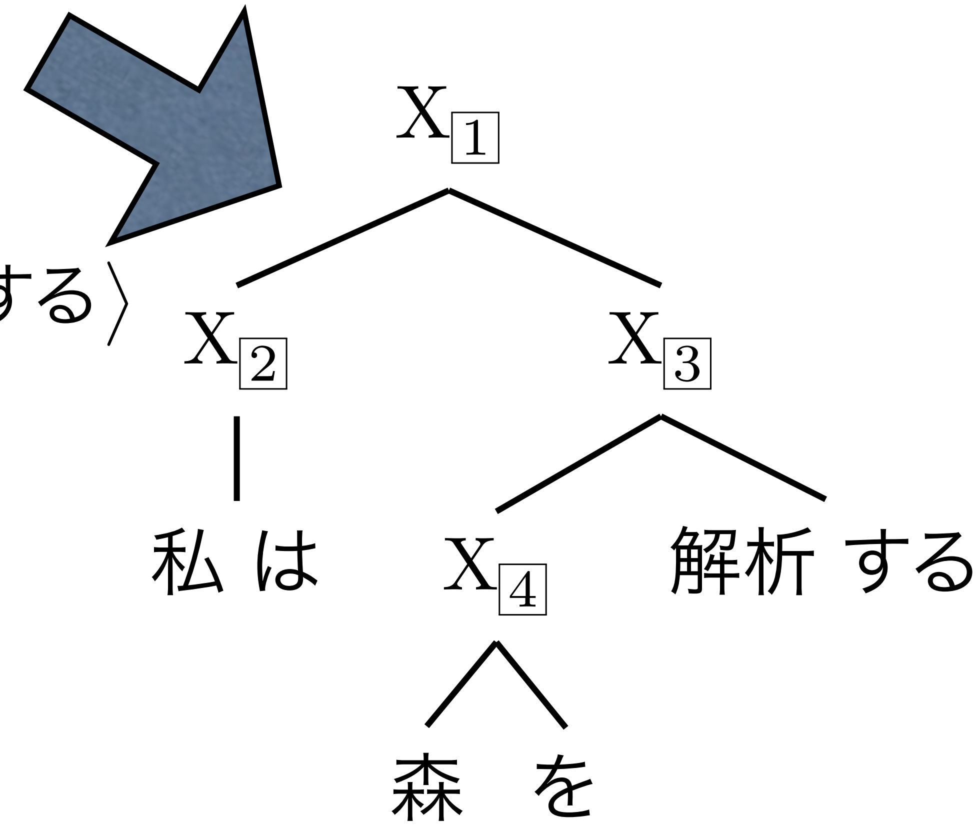


- Bitext parsing takes O(n^6) (Wu, 1997)
  - For each length n and m, for each position i and j, for each rule X ➔ LHS, for each split point k and l
- Fast span pruning by O(n^3) (Zhang et al., 2008)

Bitext parsing: two-parse



$$
\begin{aligned}
X_{\boxed{1}} &\rightarrow \langle X_{\boxed{2}}\ X_{\boxed{3}}, X_{\boxed{2}}\ X_{\boxed{3}} \rangle \\
X_{\boxed{2}} &\rightarrow \langle \text{I}, \text{私 は} \rangle \\
X_{\boxed{3}} &\rightarrow \langle \text{parse } X_{\boxed{4}}, X_{\boxed{4}} \text{ 解析 する} \rangle \\
X_{\boxed{4}} &\rightarrow \langle \text{the forest}, \text{森 を} \rangle
\end{aligned}
$$



(Dyer, 2010)

- Parse source side (Intersect with source side)
- Extract target rules from forest (relabel category)
- Parse target side by extracted rules (Compose with target side)
- The same worst case O(n^6), but fast in practice

Summary

- We reviewed some backgrounds on CFG
- Tree based MT are formulated as
  - synchronous-CFG or tree-rewrite system
  - Cube pruning allows parsing with non-local features (ngrams)

Software

- Synchronous-CFG
  - Cdec: http://cdec-decoder.org
  - Jane: http://www-i6.informatik.rwth-aachen.de/jane/
  - Joshua: http://joshua.sourceforge.net
  - Moses: http://www.statmt.org/moses/
- {Tree,String}-to-{tree, string}
  - Tiburon: http://www.isi.edu/licensed-sw/tiburon/

References

- D. Chiang, ``Hierarchical phrase-based translation," *Comput. Linguist.*, vol. 33, no. 2, pp. 201--228, 2007.
- D. Wu, ``Stochastic inversion transduction grammars and bilingual parsing of parallel corpora," *Comput. Linguist.*, vol. 23, no. 3, pp. 377--403, 1997.
- S. M. Shieber, Y. Schabes, and O. C. N. Pereira, ``Principles and implementation of deductive parsing," *Journal of Logic Programming*, 1995.
- D. Klein and C. D. Manning, ``Parsing and hypergraphs," in *In IWPT*, pp. 123--134, 2001.
- J. Goodman, ``Semiring parsing," *Comput. Linguist.*, vol. 25, no. 4, pp. 573--605, 1999.
- M. Galley, M. Hopkins, K. Knight, and D. Marcu, ``What's in a translation rule?," in *HLT-NAACL 2004: Main Proceedings* (D. M. Susan Dumais and S. Roukos, eds.), (Boston, Massachusetts, USA), pp. 273--280, Association for Computational Linguistics, May 2 - May 7 2004.
- M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang, and I. Thayer, ``Scalable inference and training of context-rich syntactic translation models," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, (Sydney, Australia), pp. 961--968, Association for Computational Linguistics, July 2006.

References

- A. Zollmann and A. Venugopal, ``Syntax augmented machine translation via chart parsing," in *StatMT '06: Proceedings of the Workshop on Statistical Machine Translation*, (Morristown, NJ, USA), pp. 138--141, Association for Computational Linguistics, 2006.
- M. Zhang, H. Jiang, A. Aw, H. Li, C. L. Tan, and S. Li, ``A tree sequence alignment-based tree-to-tree translation model," in *Proceedings of ACL-08: HLT*, (Columbus, Ohio), pp. 559--567, Association for Computational Linguistics, June 2008.
- L. Shen, J. Xu, and R. Weischedel, ``A new string-to-dependency machine translation algorithm with a target dependency language model," in *Proceedings of ACL-08: HLT*, (Columbus, Ohio), pp. 577--585, Association for Computational Linguistics, June 2008.
- Y. Ding and M. Palmer, ``Machine translation using probabilistic synchronous dependency insertion grammars," in *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, (Morristown, NJ, USA), pp. 541--548, Association for Computational Linguistics, 2005.
- Y. Liu, Y. Lu ̈, and Q. Liu, ``Improving tree-to-tree translation with packed forests," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, (Suntec, Singapore), pp. 558--566, Association for Computational Linguistics, August 2009.
- L. Huang, K. Knight, and A. Joshi, ``Statistical syntax-directed translation with extended domain of locality," in *In Proc. AMTA 2006*, pp. 66--73, 2006.

References

- Y. . Liu, Q. Liu, and S. Lin, ``Tree-to-string alignment template for statistical machine translation," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, (Sydney, Australia), pp. 609--616, Association for Computational Linguistics, July 2006.
- C. Quirk, A. Menezes, and C. Cherry, ``Depen-dency treelet translation: syntactically informed phrasal smt," in *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, (Morristown, NJ, USA), pp. 271--279, Association for Computational Linguistics, 2005.
- L. Huang and D. Chiang, ``Better k-best parsing," in *Proceedings of the Ninth InternationalWorkshop on ParsingTechnology*, (Vancouver, British Columbia), pp. 53--64, Association for Computational Linguistics, October 2005.
- H. Zhang, C. Quirk, R. C. Moore, and D. Gildea, ``Bayesian learning of non-compositional phrases with synchronous parsing," in *Proceedings of ACL-08: HLT*, (Columbus, Ohio), pp. 97--105, Association for Computational Linguistics, June 2008.
- C. Dyer, ``Two monolingual parses are better than one (synchronous parse)," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, (Los Angeles, California), pp. 263--266, Association for Computational Linguistics, June 2010.

Advanced Topics

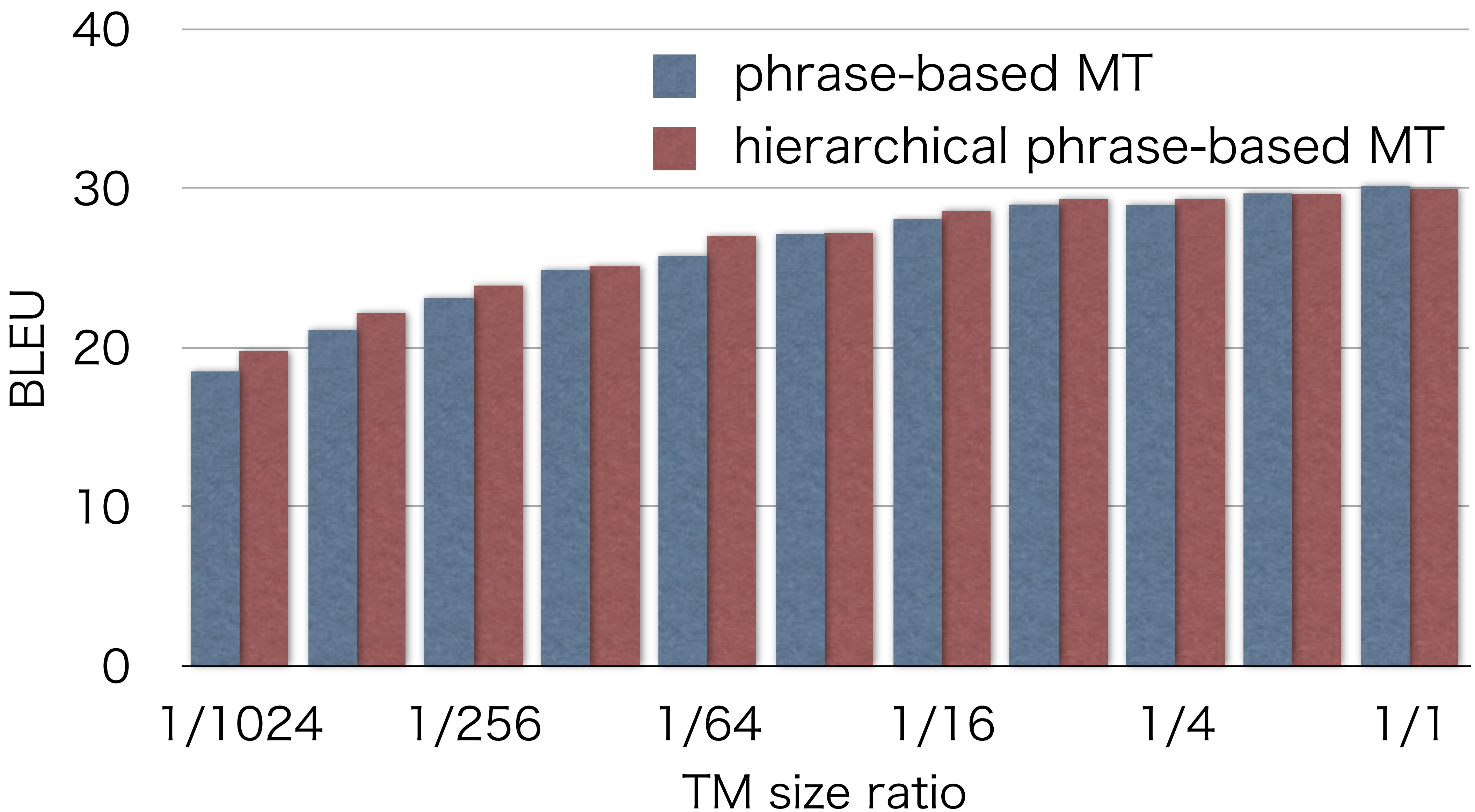
Overview

- More data, better translation?
- Translation by many features
- Single path/derivation to lattice/forest
- Word alignment, phrases, rules

More data, better translation?

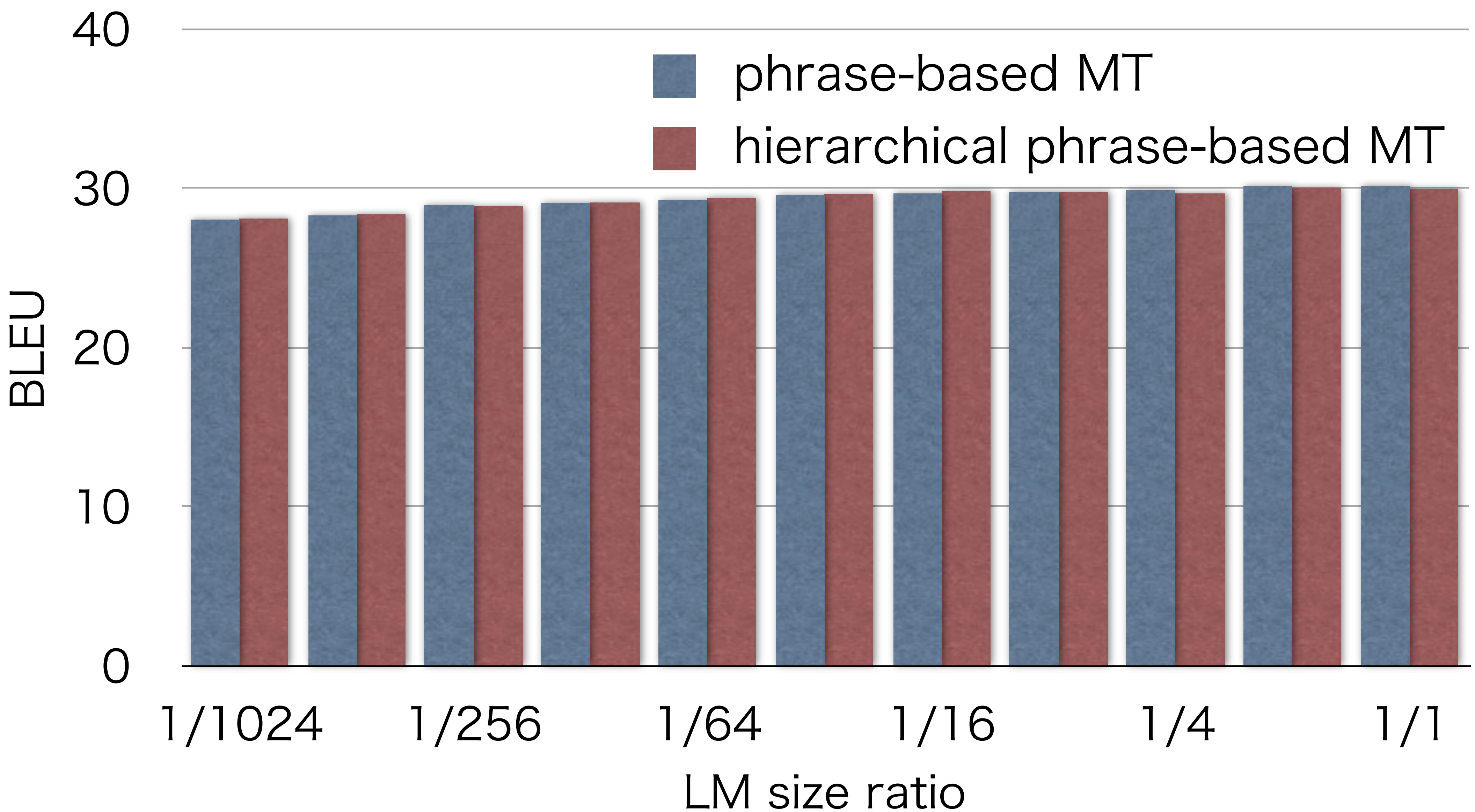
- Do we really need more data?
- Experiments on Japanese-to-English patent data
    - Language model: 11G words
    - Translation model: 108M words

Experiments: Fixed LM



- Fixed LM (11G words, 5-grams), reduced TM data (108M words)

Experiments: Fixed TM



- Fixed TM (108M words), reduced LM data (11G words)

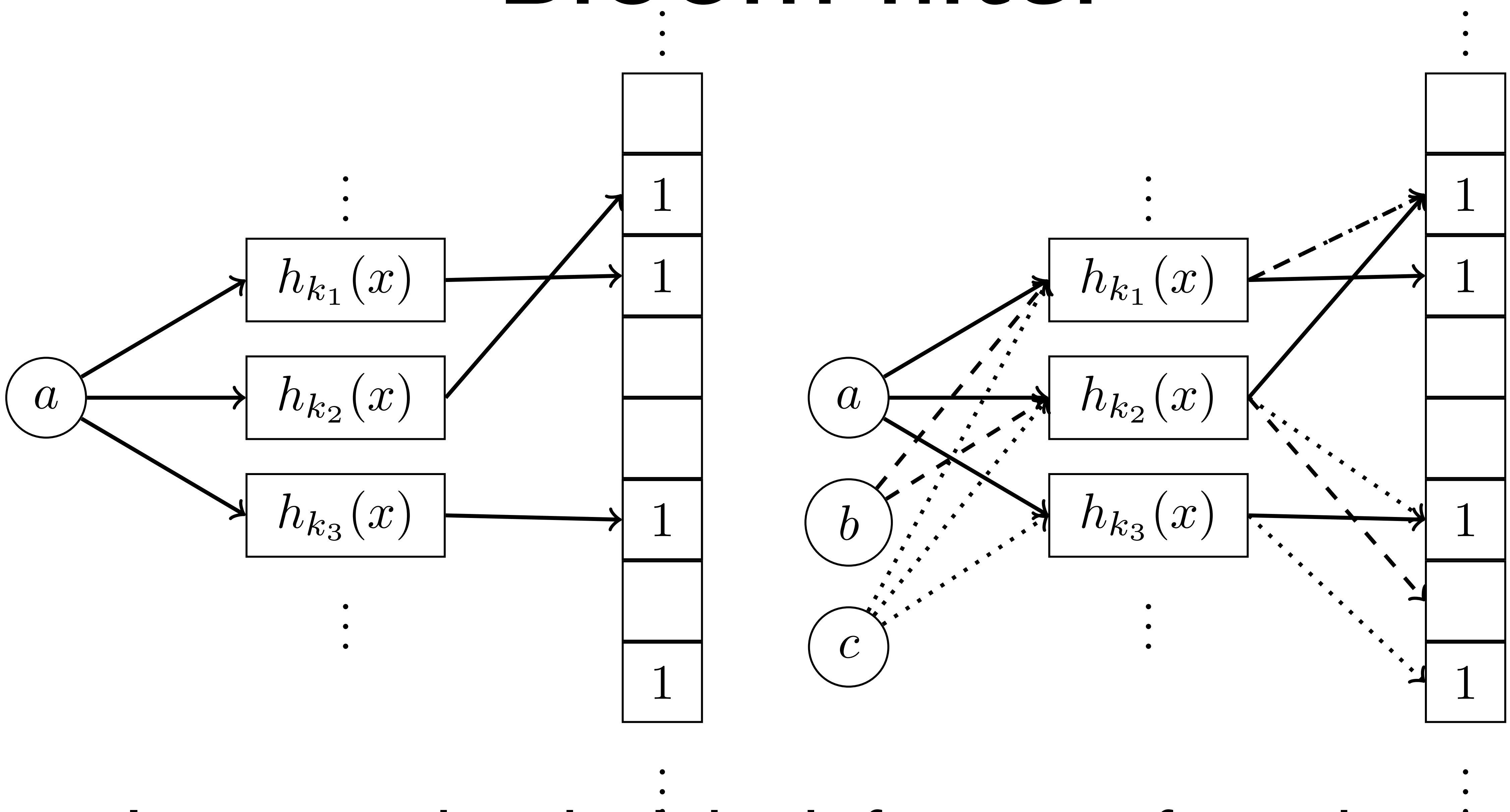
Data handling

- Parallelization (Zhang et al., 2006; Brantz et al., 2007)
  - Split data and store in clusters
  - Efficient protocol to retrieve data
- Suffix arrays (Callison-burch and Bannard, 2005; Zhang and Vogel, 2005; Lopez, 2007)
  - raw data + index by suffix array + on-the-fly phrase/rule extraction
- Alternative solutions?
  - Randomized data structures
  - Succinct data structures

Randomized data structures

- We do not store exactly, but keep signatures (Bloom, 1970)
- Allow "false positives"
  - Not inserted, but the signature says, "exists"
- Error rate is bounded theoretically and practically

Bloom filter



- Insert: set bits by k hash functions for m bits array
- Query: test by k hash functions
- False positives are controlled by k and m

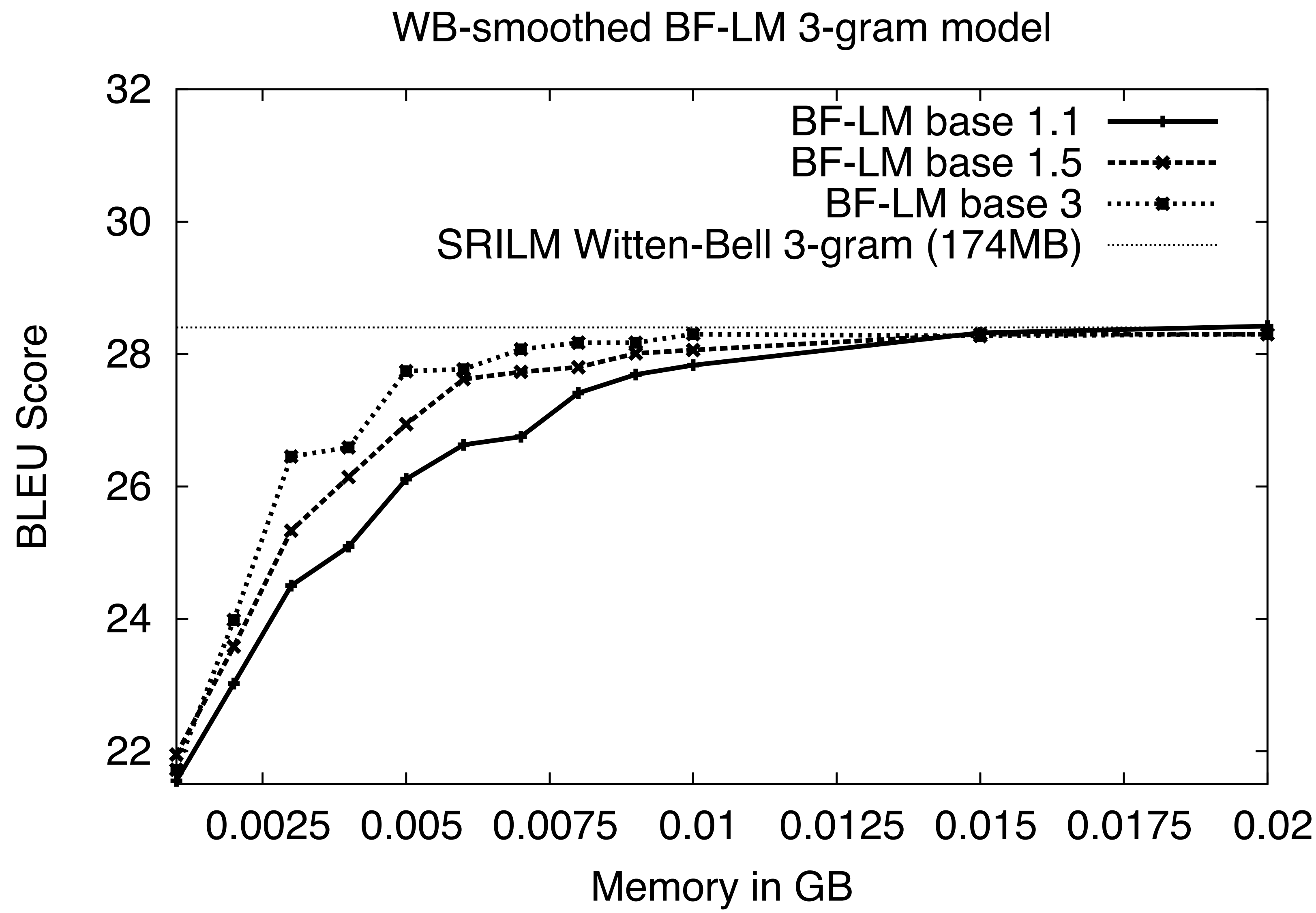
Randomized LM

for $j = 1...$ **do**
    for $i = 1...k$ **do**
        if $\mathcal{BF}\left[h_i(\{x, j\})\right] = 0$ **then**
            return $E\left[c(x) | qc(x) = j - 1\right]$
        end if
    end for
end for

(Talbot and Osborne, 2007a, 2007b)

- Store quantized log-count: $qc(x) = 1 + \lfloor \log_b c(x) \rfloor$
- Returns expected count: $E\left[c(x) | qc(x) = j\right] = \dfrac{b^{j-1} + b^j - 1}{2}$
- False positives are further controlled by ngram property:
  - If an n-gram exists, lower order (n-1)-grams exist.
  - If an n-gram exists, its count is smaller than or equal to its lower order (n-1)-grams

Randomized LM: Experiments



(Talbot and Osborne, 2007a, 2007b)

- French-English Europarl data

Other randomized variants

- Perfect hash function based randomized storage (Talbot and Brants, 2008)
- Bloomier filter which allows dynamic insertion/ deletion (Levenberg and Osborne, 2009)

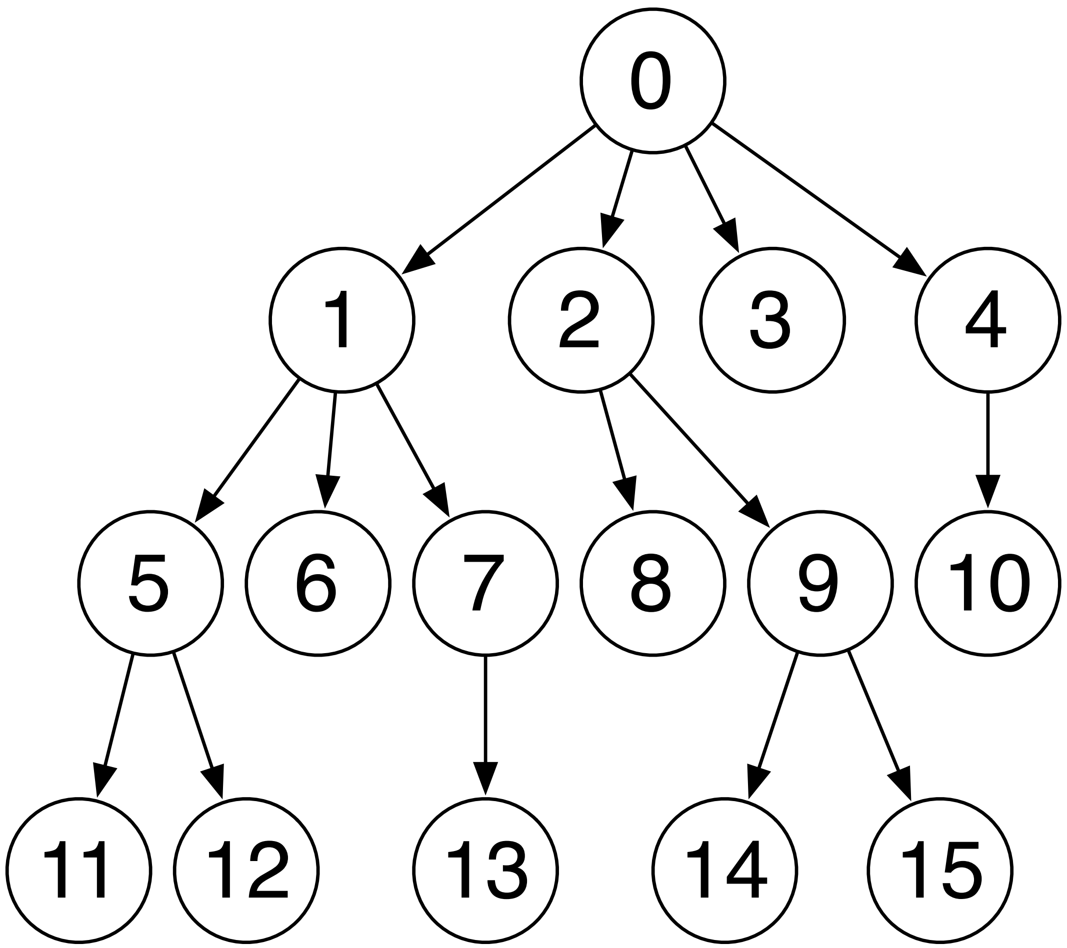
Succinct data structures

- In NLP applications (including MT), models are compactly stored by trie structures (ngrams, phrase tables, grammar etc.)
- Trie structure (pointers) can be succinctly encoded by 2M + O(M) bits, approaching information-theoretical bounds (Jacobson, 1989):

$$\lg \left\lceil \frac{1}{2M+1} \binom{2M+1}{M} \right\rceil \approx 2M - O(\lg M)$$

- An example: Level-Order Unary Degree Sequences (LOUDS) (Jacobson, 1989; Delpratt et al., 2006)

LOUDS

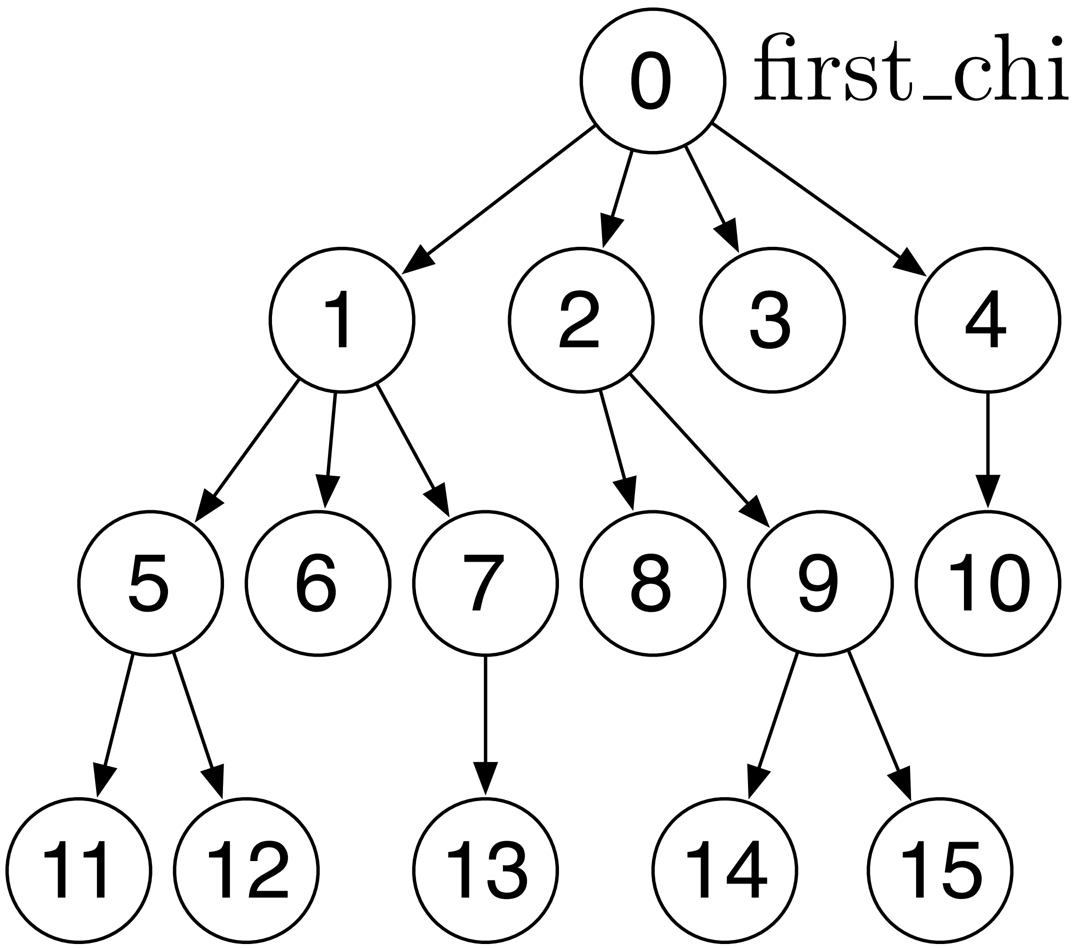


- Traverse in level order, left-to-right, emit 1s and 0 at each node
- 2M + 1 bits

node id		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
bit position	0 1	2 3 4 5 6	7 8 9 10	11 12 13	14	15 16	17 18 19	20	21 22	23	24 25 26	27	28	29	30	31	32
LOUDS bit	**10**	**11110**	**1110**	**1 1 0**	**0**	**1 0**	**1 1 0**	**0**	**1 0**	**0**	**1 1 0**	**0**	**0**	**0**	**0**	**0**	**0**

LOUDS: traversal

$$\begin{aligned}\mathrm{parent}(x) &= \mathrm{rank}_0(\mathrm{select}_1(x+1)) - 1\\ \mathrm{first\_child}(x) &= \mathrm{rank}_1(\mathrm{select}_0(x+1))\end{aligned}$$



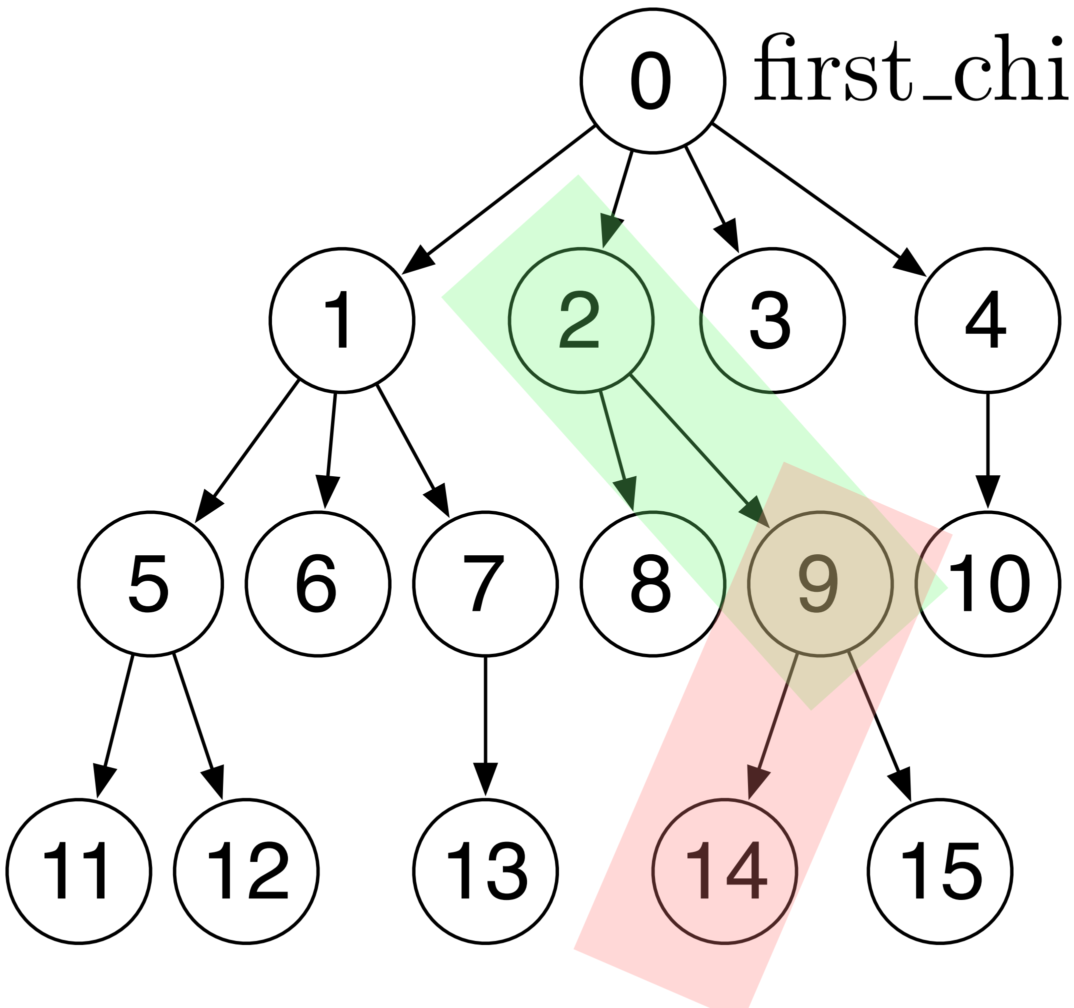
- select1(x): left-most position of the x-th bits
- rank1(x): # of bits to the left of, and including, x

node id		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
bit position	0 1	2 3 4 5 6	7 8 9 10	11 12 13	14	15 16	17 18 19	20	21 22	23	24 25 26	27	28	29	30	31	32
LOUDS bit	**10**	**11110**	**1110**	**1 1 0**	**0**	**1 0**	**1 1 0**	**0**	**1 0**	**0**	**1 1 0**	**0**	**0**	**0**	**0**	**0**	**0**

LOUDS: traversal

$$\text{parent}(x) = \text{rank}_0(\text{select}_1(x+1)) - 1$$

$$\text{first\_child}(x) = \text{rank}_1(\text{select}_0(x+1))$$



- **parent(9):**

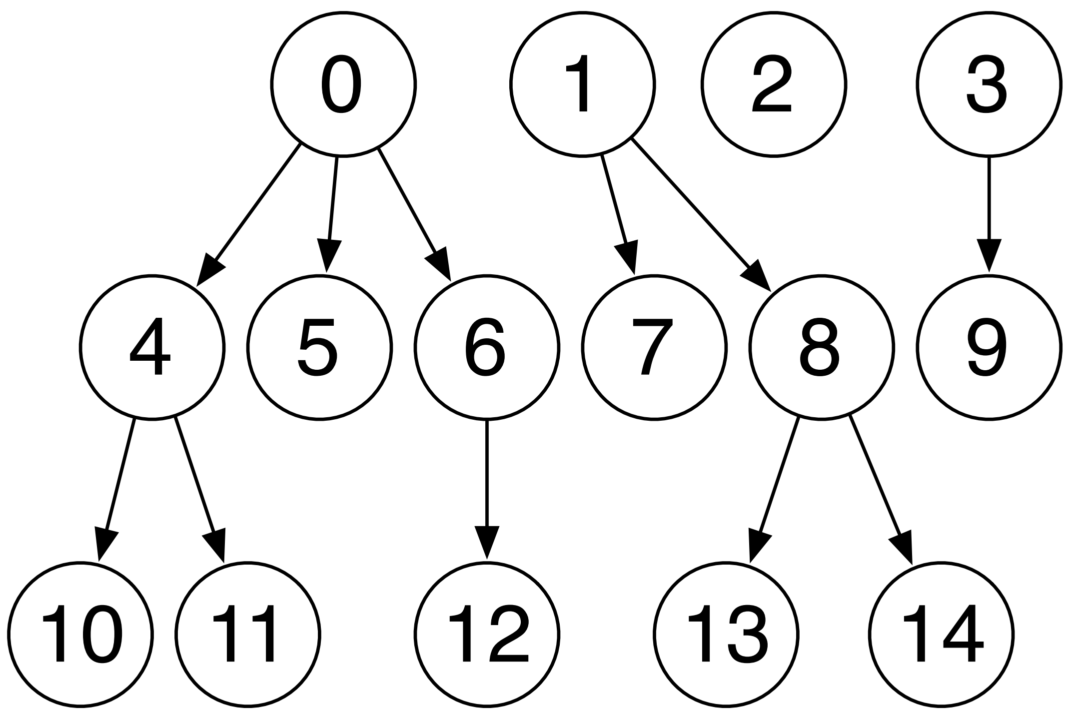
$$\text{select}_1(9+1) = 12$$
$$\text{rank}_0(12) - 1 = 2$$

- **first_child(9):**

$$\text{select}_0(9+1) = 23$$
$$\text{rank}_1(23) = 14$$

node id		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
bit position	0 1	2 3 4 5 6	7 8 9 10	11 12 13	14	15 16	17 18 19	20	21 22	23	24 25 26	27	28	29	30	31	32
LOUDS bit	1 0	1 1 1 1 0	1 1 1 0	1 1 0	0	1 0	1 1 0	0	1 0	0	1 1 0	0	0	0	0	0	0

Succinct ngram language model



- Remove root (2 bits)
- Remove the last zeros (5 bits)
- Remove unigram bits (4 + 1 bits)

$$2\mathcal{N}_1^N + 3 \rightarrow 2\mathcal{N}_1^N - (\mathcal{N}_1 + \mathcal{N}_N)$$

node id	0	1	2	3	4	5	6	7	8	9
bit position	0 1 2 3	4 5 6	7	8 9	10 11 12	13	14 15	16	17 18 19	20
LOUDS bit	**1110**	**110**	**0**	**10**	**1 1 0**	**0**	**1 0**	**0**	**1 1 0**	**0**

(watanabe et al., 2009)

Web-1T ngrams

	English	Chinese	Japanese
gzip size	25G	25G	30G
counts	12.6G	13.2	9.8G
quantized-lm	13.1G	13.8G	10.7G

- Web 1T ngrams from Google (Chinese, English, Japanese)

Software

- Randomized LM
  - randlm: http://sourceforge.net/projects/randlm/
- (generic) succinct storage
  - tx: http://code.google.com/p/tx-trie/
  - taiju: http://code.google.com/p/taiju/

Overview

- More data, better translation?
- Translation by many features
- Single path/derivation to lattice/forest
- Word alignment, phrases, rules

Model with many features

- We want fine-grained translations
- Many binary features to represent complex decision
- MERT can handle small # of features (around 10+)
- Can we scale to millions for better translations?

Large margin training

$$\hat{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}} \frac{\lambda}{2}||\mathbf{w}||^2 + \sum_{s=1}^{S} \max\left(l_s - \mathbf{w} \cdot \Delta\mathbf{h}_s\right)$$

$$\begin{aligned}
\hat{\mathbf{e}}_s &= \operatorname*{argmax}_{e} \mathbf{w} \cdot \mathbf{h}(\mathbf{e}, \mathbf{f}_s) \\
l_s &= l(\hat{\mathbf{e}}_s) - l(\mathbf{e}^*_s) \\
\Delta\mathbf{h}_s &= \mathbf{h}(\hat{\mathbf{e}}_s, \mathbf{f}_s) - \mathbf{h}(\mathbf{e}^*, \mathbf{f}_s)
\end{aligned}$$

- Major difference to MERT is the explicit L{1,2} regularizer and regression term
- Very slow convergence by SMO... faster algorithms?

(Averaged) Perceptron

Require: $\{(\mathbf{f}_s, \mathbf{e}_s)\}_{s=1}^{S}$

$\mathbf{w}^1 = \{0\}$
$t = 1$
for $1 \dots N$ **do**
  $s \sim \text{random}(1, S)$
  $\hat{\mathbf{e}} = \text{GEN}(\mathbf{f}_s, \mathbf{w}^{t-1})$
  if $l(\hat{\mathbf{e}}, \mathbf{e}_s) \geq 0$ **then**
    $\mathbf{w}^{t+1} = \mathbf{w}^t + \mathbf{h}(\mathbf{e}_s, \mathbf{f}_s) - \mathbf{h}(\hat{\mathbf{e}}, \mathbf{f}_s)$
    $t = t + 1$
  end if
end for
return $\mathbf{w}^t$ or $\frac{1}{N} \sum_{i=1}^{N} \mathbf{w}^j$

- Scales very well to very large data and large feature set
- Liang et al. (2006) reported good performance

MIRA

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{\lambda}{2} ||\mathbf{w}' - \mathbf{w}||^2 + \max\left(l_s - \mathbf{w}' \cdot \Delta\mathbf{h}_s\right)$$

$$\begin{aligned} \hat{\mathbf{e}}_s &= \underset{e}{\operatorname{argmax}}\, \mathbf{w} \cdot \mathbf{h}(\mathbf{e}, \mathbf{f}_s) \\ l_s &= l(\hat{\mathbf{e}}_s) - l(\mathbf{e}_s^*) \\ \Delta\mathbf{h}_s &= \mathbf{h}(\hat{\mathbf{e}}_s, \mathbf{f}_s) - \mathbf{h}(\mathbf{e}^*, \mathbf{f}_s) \end{aligned}$$

- line 7 of weight update is replaced by the solution of the above equation
- Similar to large margin constraints
- Experimented by: Watanabe et al. (2007); Chiang et al. (2008); Chiang et al. (2009)

Correct translations?

$$
\begin{aligned}
\hat{\mathbf{e}}_s &= \operatorname*{argmax}_e \mathbf{w} \cdot \mathbf{h}(\mathbf{e}, \mathbf{f}_s) - \mathrm{BLEU}_s(\mathbf{e}) \\
\mathbf{e}^*_s &= \operatorname*{argmax}_e \mathbf{w} \cdot \mathbf{h}(\mathbf{e}, \mathbf{f}_s) + \mathrm{BLEU}_s(\mathbf{e})
\end{aligned}
$$

- Problem: we cannot generate translations exactly the same as reference translations.
- Solution: select translations among nbests with "error bias" (Chiang et al., 2008; Chian et al., 2009)

MIRA: Experiments

System	Training	Features	#	Tune	Test
Hiero	MERT	baseline	11	35.4	36.1
	MIRA	syntax, distortion	56	35.9	36.9*
		syntax, distortion, discount	61	36.6	37.3**
		all source-side, discount	10990	38.4	37.6**
Syntax	MERT	baseline	25	38.6	39.5
	MIRA	baseline	25	38.5	39.8*
		overlap	132	38.7	39.9*
		node count	136	38.7	40.0**
		all target-side, discount	283	39.6	40.6**

(Chiang et al., 2009)

- Consistent improvements over MERT
- Scales well to millions of features (Watanabe et al., 2007)

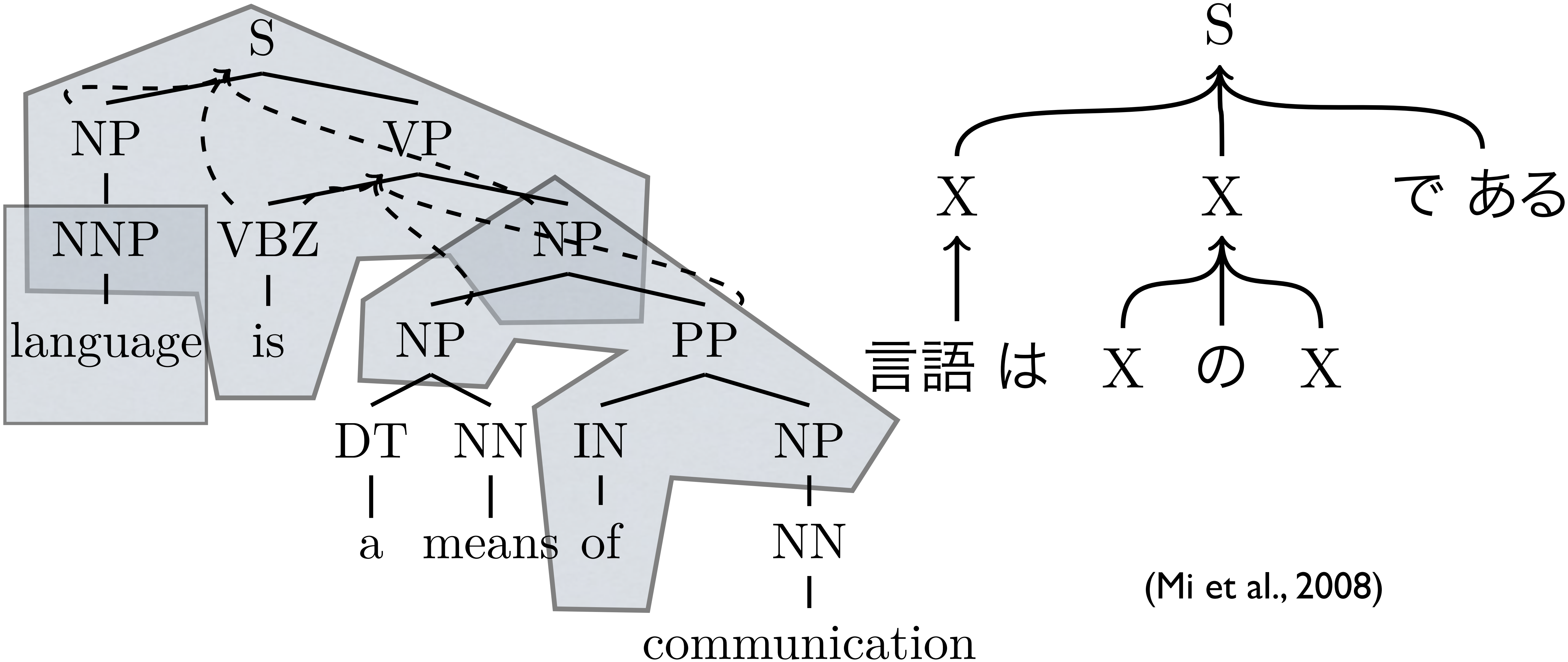
Overview

- More data, better translation?
- Translation by many features
- Single path/derivation to lattice/forest
- Word alignment, phrases, rules

Forest approaches

- single {tree, string} input and single {tree,s tring} output
- As in lattice/word graph, we can compactly represent alternative derivations by forest
- Translation from forest, Extraction from forest, MBR by forest, MERT by forest

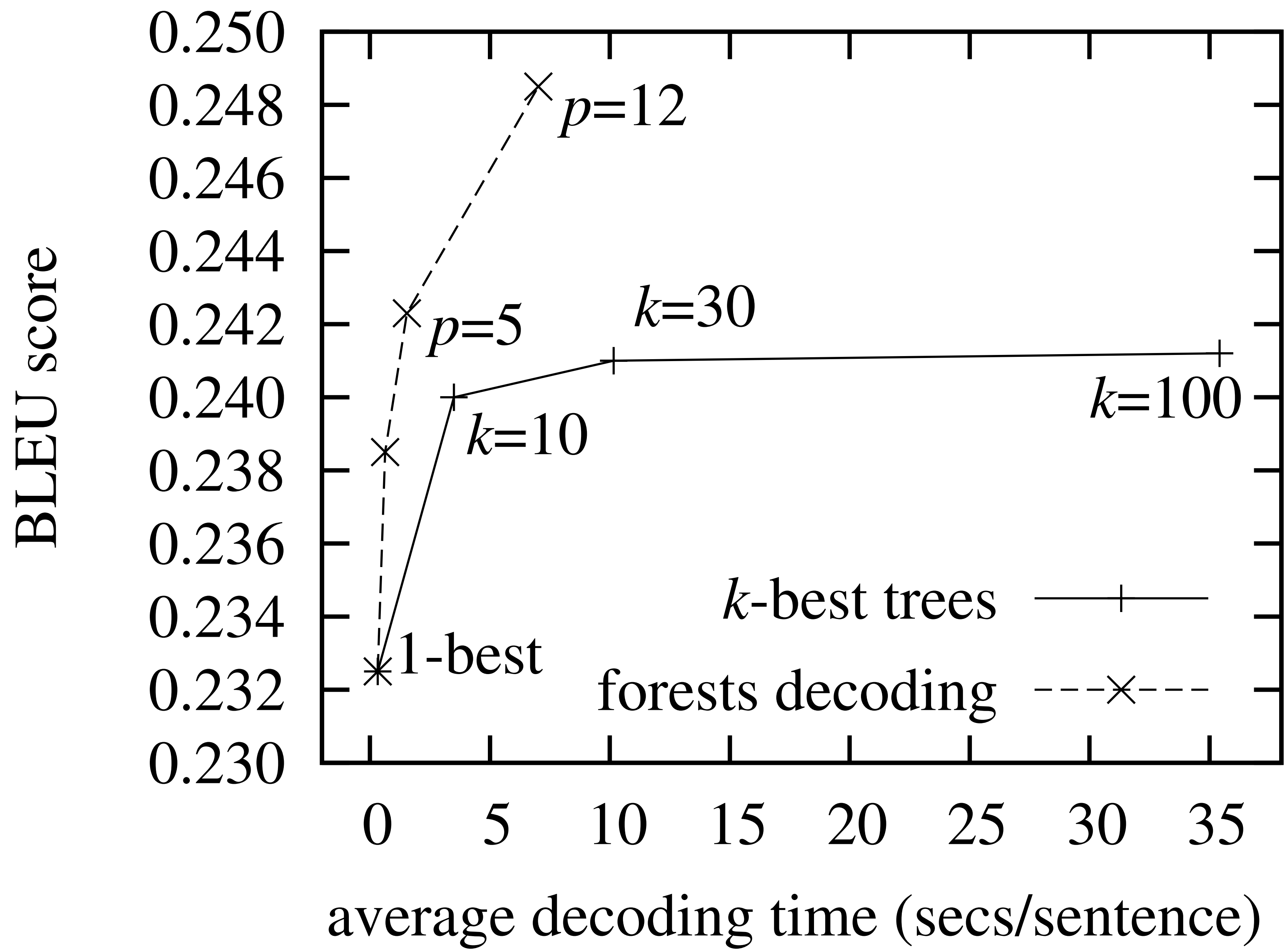
Translation from forest



(Mi et al., 2008)

- (Try) avoid errors propagated from parse tree, and make decision later
- Tree rewrite on forest, yielding larger translation forest

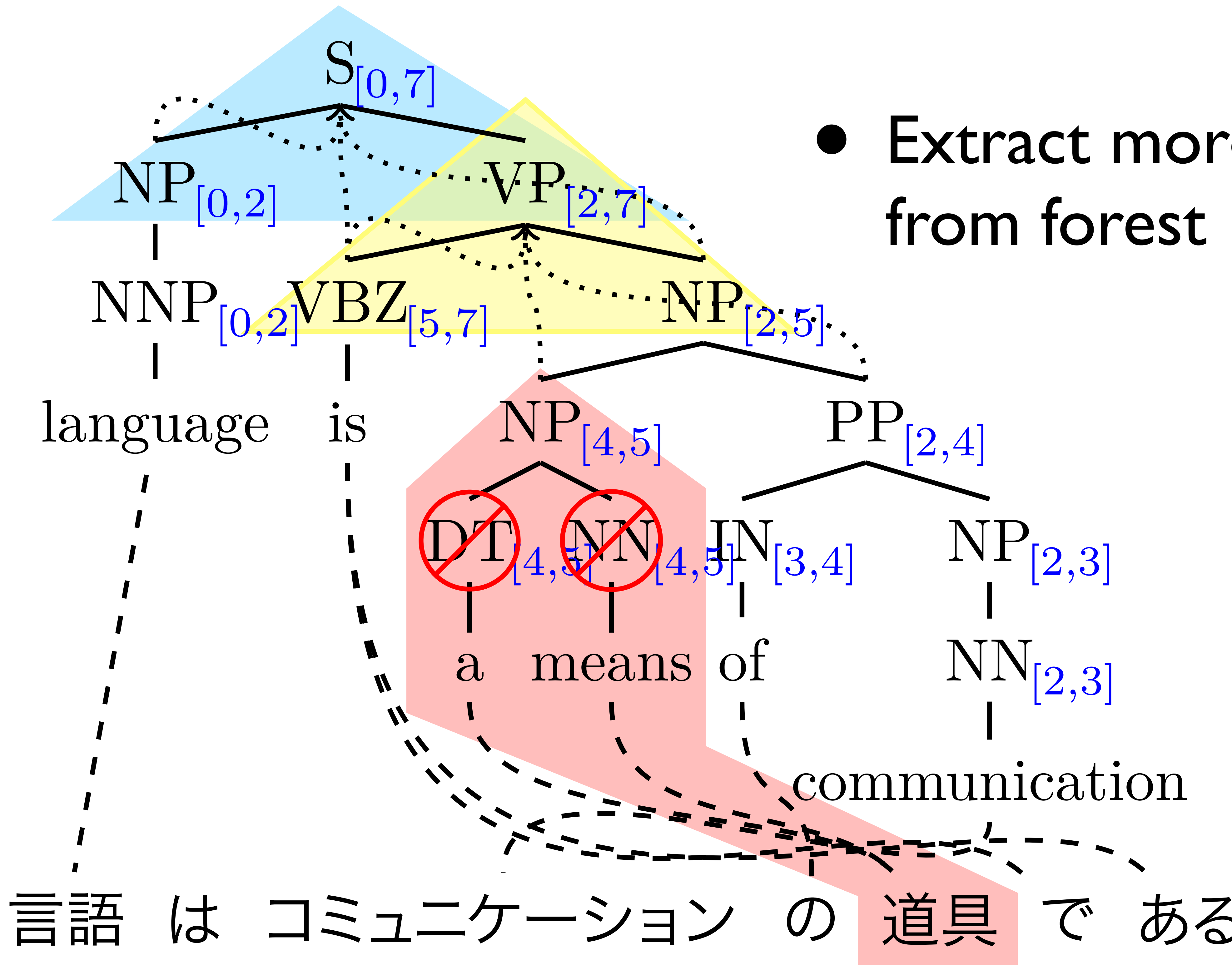
Translation from forest



(Mi et al., 2008)

- Faster than translating each of k-best trees
- Better translations from packed forest

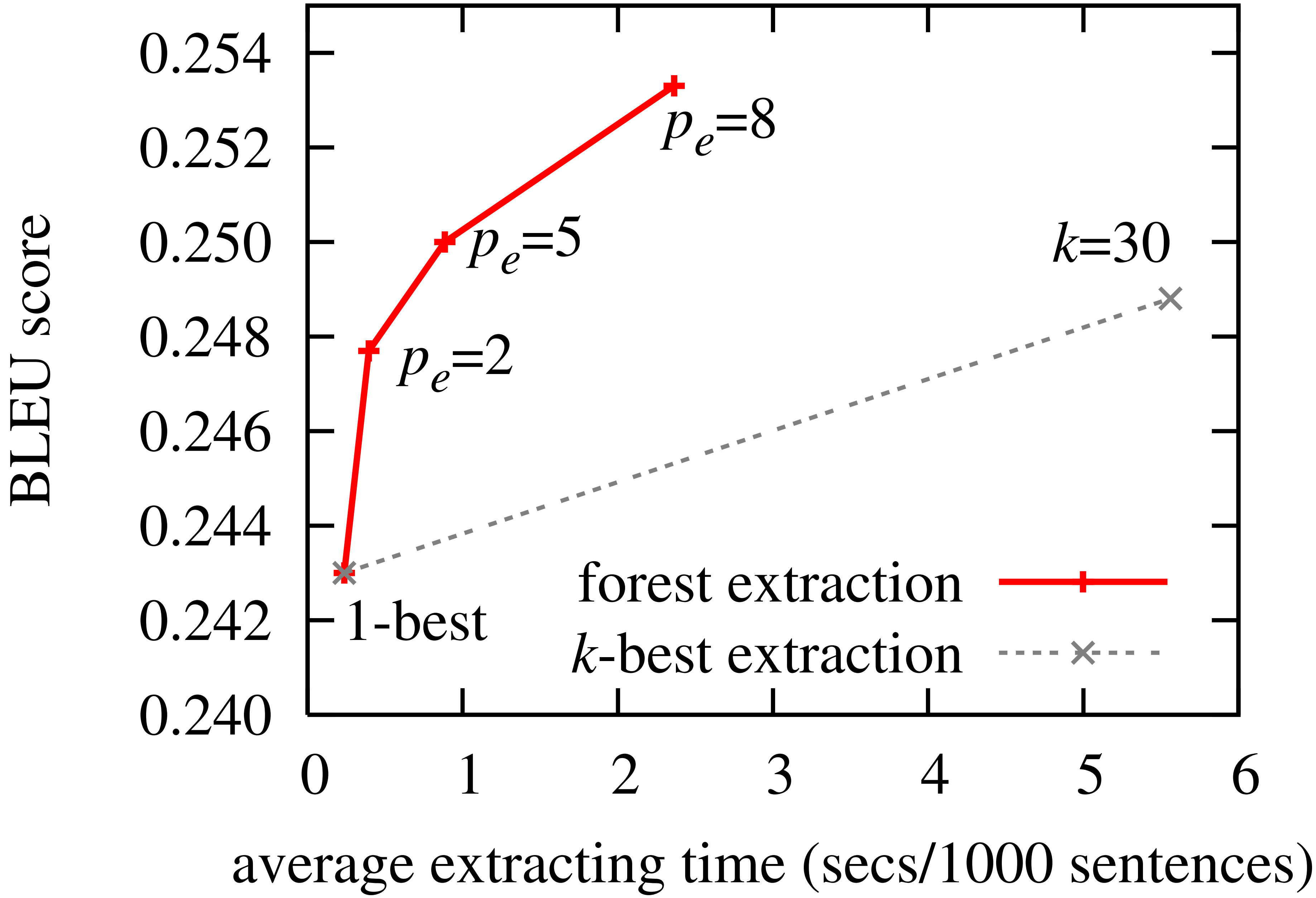
Extraction from forest



- Extract more rules from forest

(Mi and Huang, 2008)

Extraction from forest



- Faster than extraction from individual trees
- Better translations from larger forest

MBR by forest

$$\begin{aligned}\hat{\mathrm{e}} &= \operatorname*{argmin}_{\mathbf{e}} \mathbb{E}_{P(\mathbf{e}'|\mathbf{f})}\left[l(\mathbf{e};\mathbf{e}')\right] \\ &= \operatorname*{argmin}_{\mathbf{e}} \sum_{e'} l(\mathbf{e};\mathbf{e}')P(\mathbf{e}'|\mathbf{f})\end{aligned}$$

- Instead of maximization, we reduce expected loss (MBR, Minimum Bayes Risk)
- Conventional approaches enumerate over n-best-list (Kumar and Byrne, 2004)

MBR by linear BLEU

$$
\begin{aligned}
l(\mathbf{e};\mathbf{e}') &= \theta_0|\mathbf{e}| + \sum_{w \in N} \theta_{|w|} c_w(e) \delta_w(e') \\
\hat{e} &= \operatorname*{argmax}_{\mathbf{e} \in \mathcal{G}} \theta_0|\mathbf{e}| + \sum_{w} \theta_{|w|} c_w(e) p(w|\mathcal{G})
\end{aligned}
$$

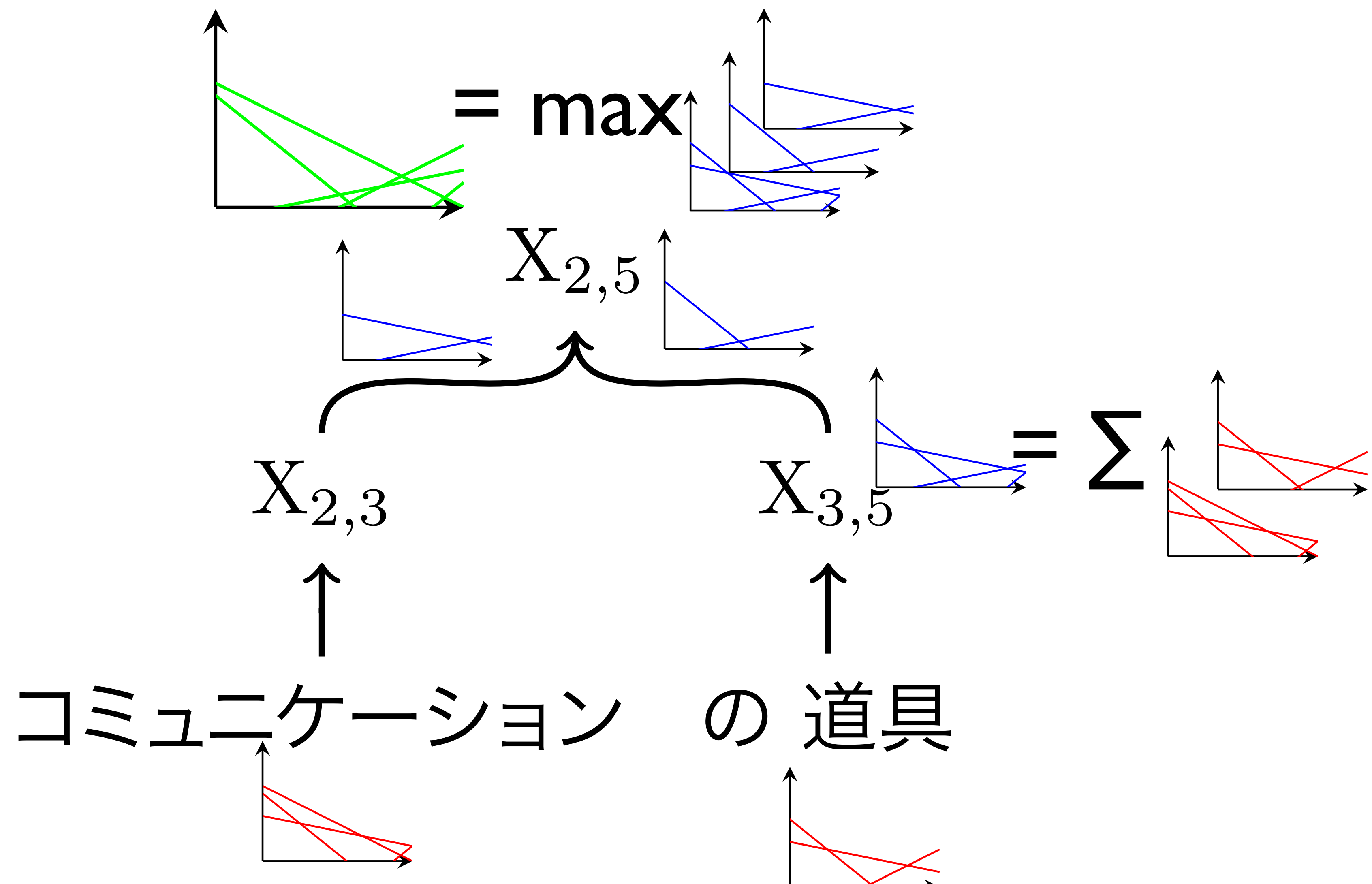
- When computing expected loss (= 1.0 - BLEU) over lattice/forest, use linearly approximated BLEU (Tromble et al., 2008, Kumar et al., 2009)

MBR by expected BLEU

$$
\begin{aligned}
BLEU(\mathbf{e};\mathbf{e}') &= \exp\left(\min(1-\frac{|\mathbf{e}'|}{|\mathbf{e}|}) + \frac{1}{4}\sum_{n=1}^{4}\log p_n(\mathbf{e},\mathbf{e}')\right) \\
p_n(\mathbf{e},\mathbf{e}') &= \frac{\sum_{w\in\mathcal{T},|w|=n}\min(c(\mathbf{e},w),c(\mathbf{e}',w))}{\sum_{w\in\mathcal{T},|w|=n}c(\mathbf{e},w)}
\end{aligned}
$$

- As an alternative to MBR, compute similarities by expected ngram statistics (DeNero et al., 2009)
- expected ngram counts for e' are collected from hypergraph T

MERT by forest



(Kumar et al., 2009)

- MERT is performed over forest, not n-best
- Hyperedge: combine lines from antecedents
- Node: Compute convex hulls for maximization

Overview

- More data, better translation?
- Translation by many features
- Single path/derivation to lattice/forest
- Word alignment, phrases, rules

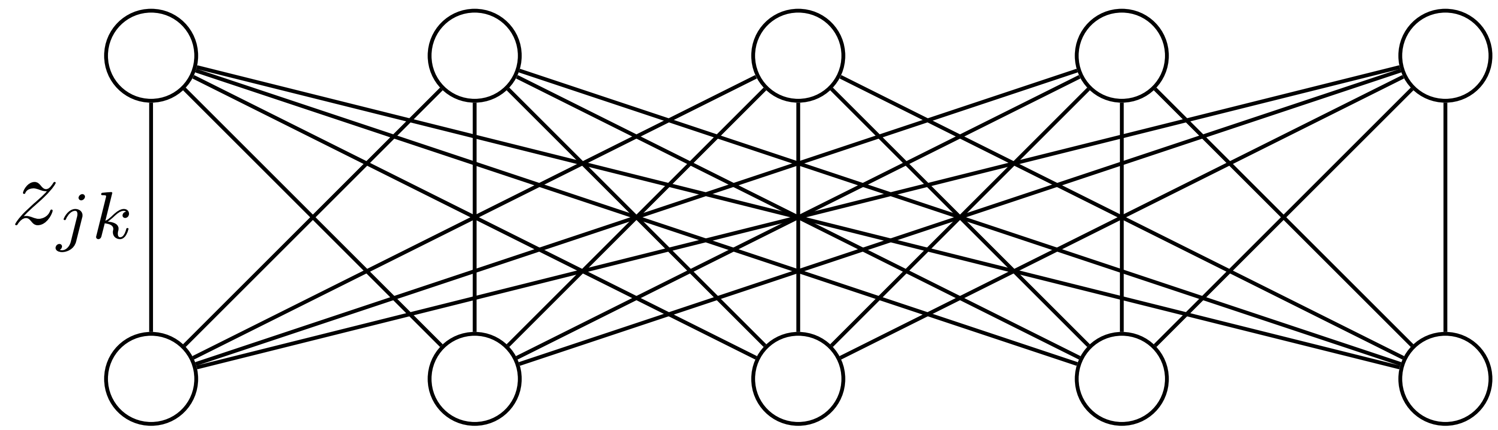
Word alignment, phrases, rules

- Better word alignment learning?
  - We learned “unsupervised” word alignment training
  - What if “gold standard” exists?
- Better phrases, rules?
  - We can extract phrases/rules from word alignment annotated data
  - Can we directly induce phrases/rules?

Supervised word alignment

- IBM Models and HMM model can learn from bilingual sentences
- No control on “how word will be aligned”
- Assuming small data with word alignment annotation
- max-matching, ITG, Block-ITG, ITG+bi-parse

Max-matching alignment



$$
\begin{aligned}
\max_{\mathbf{z}} \quad & \sum_{j,k} s_{jk} z_{jk} \\
\text{s.t.} \quad & \sum_{j} z_{jk} \leq 1, \sum_{k} z_{jk} \leq 1, 0 \leq z_{jk} \leq 1 \\
& s_{jk} = \mathbf{w} \cdot \mathbf{h}(\mathbf{e}_j, \mathbf{f}_k)
\end{aligned}
$$

- Word alignment as a max-flow problem over bipartite graph (Taskar et al., 2005)
- Solved by the linear program
- Max-margin training for parameter estimation

ITG alignment

$$
\begin{array}{lllllll}
\mathrm{X} & \rightarrow & [\mathrm{X}\ \mathrm{X}] & \quad & \mathrm{X} & \rightarrow & \langle \mathrm{X}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}}, \mathrm{X}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}} \rangle \\
\mathrm{X} & \rightarrow & \langle \mathrm{X}\ \mathrm{X} \rangle & \quad & \mathrm{X} & \rightarrow & \langle \mathrm{X}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}}, \mathrm{X}_{\boxed{2}}\ \mathrm{X}_{\boxed{1}} \rangle \\
\mathrm{X} & \rightarrow & e/f & \quad & \mathrm{X} & \rightarrow & \langle e, f \rangle
\end{array}
$$

- Binary branching rules
- non-ambiguous deletion by Haghighi et al. (2009)
- Leraning by EM-algorithm (Wu, 1997), or, max-margin training (Cherry and Lin, 2006)

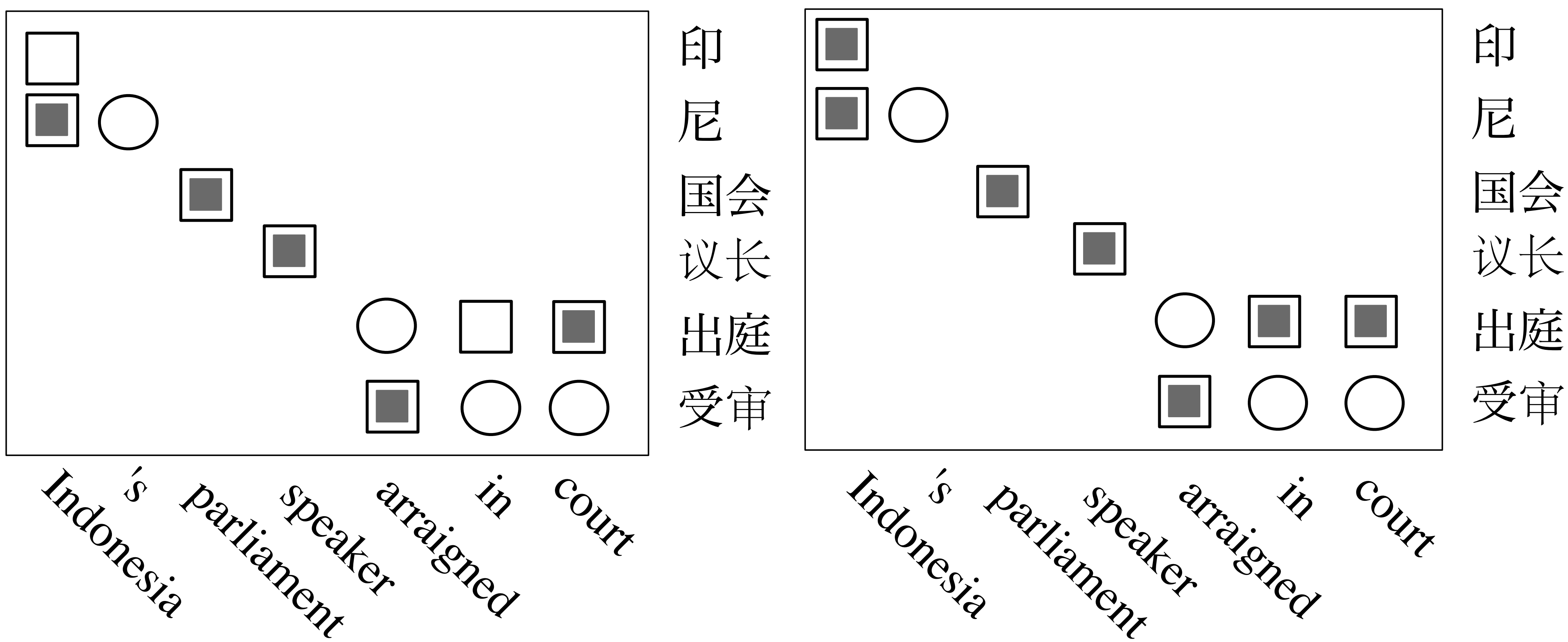
ITG-alignment: Experiments

Method	Prec	Rec	AER
Matching	0.916	0.860	0.110
D-ITG	0.940	0.854	0.100
SD-ITG	0.944	0.878	0.086

(Cherry and Lin, 2006)

- Experiments with dependency constraint
- Evaluated by alignment error rate (AER)
- Still, it is not clear whether improved alignment implies improved BLEU

Block ITG-alignment



- Allow phrasal alignment by adding phrasal lexical rules (Haghighi et al., 2009)

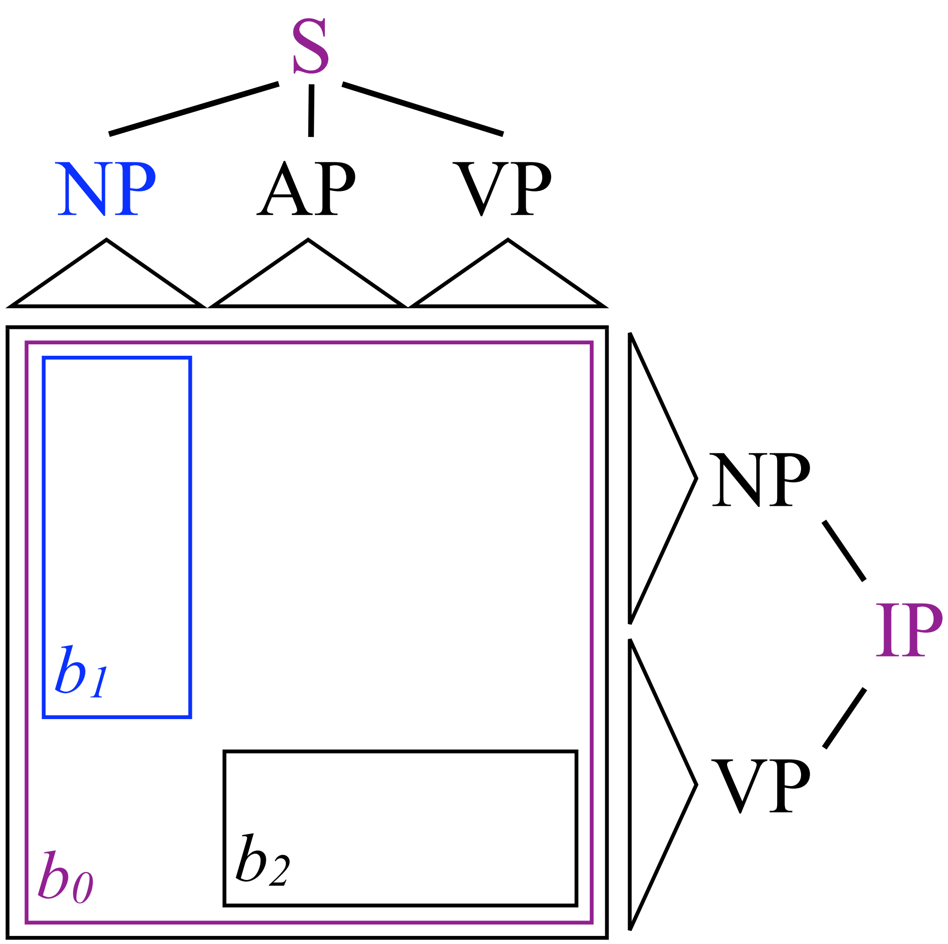
Block ITG-alignment: Experiments

Alignments			Translations	
Model	Prec	Rec	Rules	BLEU
GIZA++	62	84	1.9M	23.22
Joint HMM	79	77	4.0M	23.05
Viterbi ITG	90	80	3.8M	24.28
Posterior ITG	81	83	4.2M	**24.32**

(Haghighi et al., 2009)

- Chinese/English translation
- Large margin-based MIRA training and MaxEnt traning
- The first work to show gain by alignment improved BLEU

ITG + Bi-parsing alignment



Features	
ϕ (IP, s)	ϕ (b_0, s, s')
ϕ (NP, s)	ϕ (b_1, s, s')
ϕ (VP, s)	ϕ (b_2, s, s')
ϕ (S, s')	$\phi_{\triangleright}$ (IP, b_0)
ϕ (NP, s')	$\phi_{\triangleleft}$ (b_0, S)
ϕ (AP, s')	$\phi_{\triangleleft}$ (b_1, NP)
ϕ (VP, s')	$\phi_{\bowtie}$ (IP, b_0, S)

Parsing

Alignment

Synchronization

(Burkett et al., 2010)

- ITG-alignment with syntactic parses from source/target
- Asynchronous features: no direct pairing features
- Mean field inference for approximate estimation

ITG + Bi-parsing alignment

	Test Results			
	Precision	Recall	AER	F_1
HMM	86.0	58.4	30.0	69.5
ITG	**86.8**	73.4	20.2	79.5
Joint	85.5	**84.6**	**14.9**	**85.0**

	Rules	Tune	Test
HMM	1.1M	29.0	29.4
ITG	1.5M	**29.9**	30.4†
Joint	1.5M	29.6	**30.6**

(Burkett et al., 2010)

- Gain from Haghighi et al. (2009)

Direct phrase/rule induction

- We have separated word alignment and phrase/rule induction
- Can we learn directly?

Direct phrase training

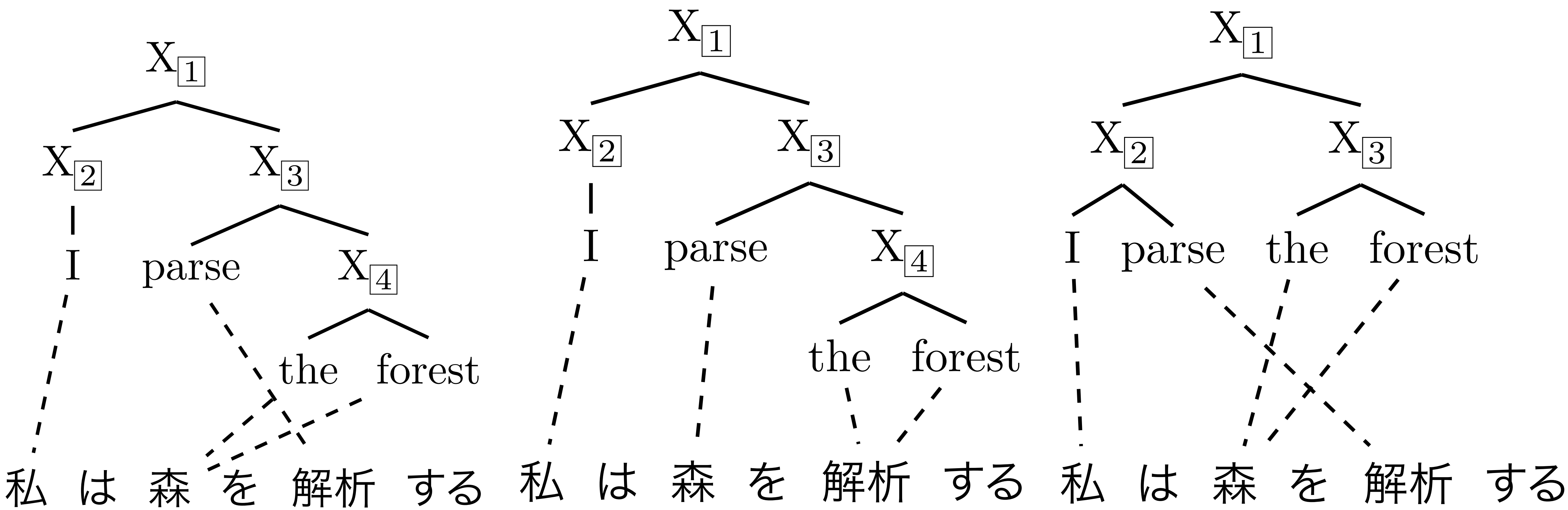
- Instead of training from word alignment data, why not directly train phrases, rules?
- Many work: Marcu and Wong (2002) etc.
- Some of the problems:
    - Very expensive summation
    - EM-algorithm w/o control by prior belief: use of non-parametric Bayesian approach

Optimization/Summation

	optimization	summation
tractable	A*/Knuth/Viterbi	forward-backward/ inside-outside
intractable	beam search	???

- We need summation for training parameters
  - Margin-based or Loss-based learning avoid this problem
- DP-based algorithm is applicable to tractable models
- Our choice: tractable simpler (and often approximated) model or complex model w/o approximation?

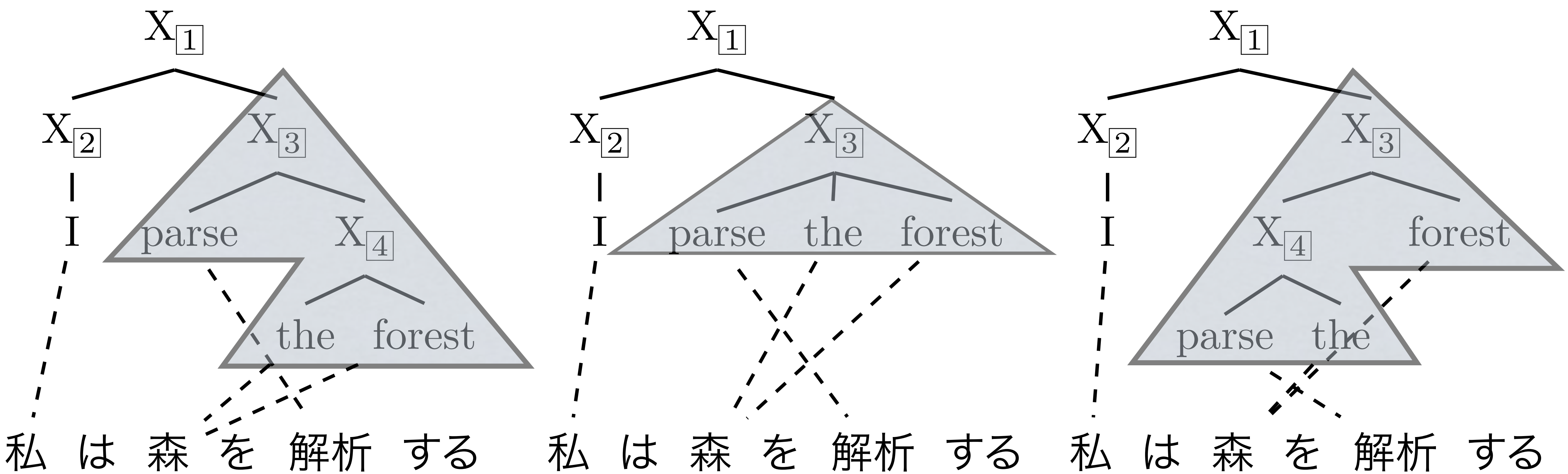
Monte Carlo algorithms



$$p(\mathrm{Y} = \{\mathrm{tree}\ ,\ \mathrm{alignment}\} | \mathrm{X} = \{\mathrm{I\ parse\ ...,\ 私\ は\ ...}\})$$

- Instead of DP based summing, sampling

Markov Chain Monte Carlo



$P(Y_1|Y_0)$ $P(Y_2|Y_1)$ $P(Y_3|Y_2)$

- Sampling by a series of small changes

Summation problem: Summary

- MCMC for intractable models
- Define your sampling operations
- Examples:
  - Phrase-based models (DeNero et al., 2008; Arun et al., 2009)
  - Synchronous-CFG (Blunsom et al., 2009)
  - string-to-tree (Cohn and Blunsom, 2009)

MCMC: efficient samplings

- Block sampling (Cohn and Blunsom, 2010):
  - Allow larger changes by simultaneously perform small changes
- Slice sampling (Blunsom and Cohn, 2010):
  - Together with block sampling, pruning parameter determined by model
- Randomized pruning (Bouchard-Coˆte ´ et al., 2009):
  - Sampling over “invalid spans” instead of trees

Summary

- Promising direction by nonparametric Bayesian approaches
- Sampling methods replace DP-based training
- Alternative: Variational approaches inspired by DP-based training

Conclusion

Outlook: Progress in 20 years

- Modeling: word to phrase, tree, forest
- Search: even with complex structural modeling, we can search efficiently
- Training: large contribution from Machine Learning techniques
- Computer Science: CPU, memory, parallelization, data structure

Outlook: Future?

- More data with less structure or less data with more structures
- General translation or task-specific translation
- Your contributions!

References

- G. Jacobson, ``Space-efficient static trees and graphs," in *30th Annual Symposium on Foundations of Computer Science*, pp. 549--554, Nov 1989.
- O. Delpratt, N. Rahman, and R. Raman, ``Engineering the LOUDS succinct tree representation," in *Proceedings of the 5th International Workshop on Experimental Algorithms*, pp. 134--145, 2006.
- K. Church, T. Hart, and J. Gao, ``Compressing trigram language models with Golomb coding," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 199--207, 2007.
- D. Talbot and M. Osborne, ``Smoothed Bloom filter language models: Tera-scale LMs on the cheap," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 468--476, 2007.
- T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean, ``Large language models in machine translation," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 858--867, 2007.
- D. Talbot and T. Brants, ``Randomized language models via perfect hash functions," in *Proceedings of ACL-08: HLT*, (Columbus, Ohio), pp. 505--513, June 2008.

References

- D. Talbot and M. Osborne, ``Randomised language modelling for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, (Prague, Czech Republic), pp. 512--519, June 2007.
- Y. Zhang and S. Vogel, ``An efficient phrase-to-phrase alignment model for arbitrarily long phrase and large corpora," in *In Proceedings of the 10th Con*ference of the European Association for Machine Translation (EAMT-05), pp. 30--31, 2005.
- Y. Zhang, A. S. Hildebrand, and S. Vogel, ``Distributed language modeling for n-best list reranking," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, (Sydney, Australia), pp. 216--223, July 2006.
- C. Callison-burch and C. Bannard, ``Scaling phrase-based statistical machine translation to larger corpora and longer phrases," in *In Proceedings of ACL*, pp. 255--262, 2005.
- A. Lopez, ``Hierarchical phrase-based translation with suffix arrays," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, (Prague, Czech Republic), pp. 976--985, Association for Computational Linguistics, June 2007.
- T. Watanabe, H. Tsukada, and H. Isozaki, ``A succinct n-gram language model," in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, (Suntec, Singapore), pp. 341--344, Association for Computational Linguistics, August 2009.
- B. H. Bloom, ``Space/time trade-offs in hash coding with allowable errors," *Commun. ACM*, vol. 13, no. 7, pp. 422--426, 1970.

References

- A. Levenberg and M. Osborne, ``Stream-based randomised language models for SMT," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, (Singapore), pp. 756--764, Association for Computational Linguistics, August 2009.
- B. Taskar, S. Lacoste-Julien, and D. Klein, ``A discriminative matching approach to word alignment," in *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, (Morristown, NJ, USA), pp. 73--80, Association for Computational Linguistics, 2005.
- C. Cherry and D. Lin, ``Soft syntactic constraints for word alignment through discriminative training," in *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, (Sydney, Australia), pp. 105--112, Association for Computational Linguistics, July 2006.
- A. Haghighi, J. Blitzer, J. DeNero, and D. Klein, ``Better word alignments with supervised itg models," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, (Suntec, Singapore), pp. 923--931, Association for Computational Linguistics, August 2009.
- D. Burkett, J. Blitzer, and D. Klein, ``Joint parsing and alignment with weakly synchronized grammars," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, (Los Angeles, California), pp. 127--135, Association for Computational Linguistics, June 2010.
- D. Smith and J. Eisner, ``Quasi-synchronous grammars: Alignment by soft projection of syntactic dependencies," in *Proceedings on the Workshop on Statistical Machine Translation*, (New York City), pp. 23--30, Association for Computational Linguistics, June 2006.

References

- P. Liang, A. Bouchard-Coˆte ´, D. Klein, and B. Taskar, ``An end-to-end discriminative approach to machine translation," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, (Sydney, Australia), pp. 761--768, Association for Computational Linguistics, July 2006.
- T. Watanabe, J. Suzuki, H. Tsukada, and H. Isozaki, ``Online Large-Margin Training for Statistical Machine Translation," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, (Prague, Czech Republic), pp. 764--773, June 2007.
- D. Chiang, Y. Marton, and P. Resnik, ``Online large-margin training of syntactic and structural translation features," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, (Honolulu, Hawaii), pp. 224--233, Association for Computational Linguistics, October 2008.
- D. Chiang, K. Knight, and W. Wang, ``11,001 new features for statistical machine translation," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, (Boulder, Colorado), pp. 218--226, Association for Computational Linguistics, June 2009.
- D. Marcu and W. Wong, ``A phrase-based, joint probability model for statistical machine translation," in *Proc. of EMNLP-2002*, (Philadelphia, PA), July 2002.
- H. Mi, L. Huang, and Q. Liu, ``Forest-based translation," in *Proceedings of ACL-08: HLT*, (Columbus, Ohio), pp. 192--199, Association for Computational Linguistics, June 2008.
- H. Mi and L. Huang, ``Forest-based translation rule extraction," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, (Honolulu, Hawaii), pp. 206--214, Association for Computational Linguistics, October 2008.

References

- S. Kumar, W. Macherey, C. Dyer, and F. Och, ``Efficient minimum error rate training and minimum bayes-risk decoding for translation hypergraphs and lattices," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, (Suntec, Singapore), pp. 163--171, Association for Computational Linguistics, August 2009.
- DeNero, D. Chiang, and K. Knight, ``Fast consensus decoding over translation forests," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, (Suntec, Singapore), pp. 567--575, Association for Computational Linguistics, August 2009.
- J. DeNero, A. Bouchard-Coˆte ´, and D. Klein, ``Sampling alignment structure under a Bayesian translation model," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, (Honolulu, Hawaii), pp. 314--323, Association for Computational Linguistics, October 2008.
- A. Arun, C. Dyer, B. Haddow, P. Blunsom, A. Lopez, and P. Koehn, ``Monte carlo inference and maximization for phrase-based translation," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, (Boulder, Colorado), pp. 102--110, Association for Computational Linguistics, June 2009.
- P. Blunsom, T. Cohn, C. Dyer, and M. Osborne, ``A gibbs sampler for phrasal synchronous grammar induction," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, (Suntec, Singapore), pp. 782--790, Association for Computational Linguistics, August 2009.
- T. Cohn and P. Blunsom, ``A Bayesian model of syntax-directed tree to string grammar induction," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, (Singapore), pp. 352--361, Association for Computational Linguistics, August 2009.
- T. Cohn and P. Blunsom, ``Blocked inference in bayesian tree substitution grammars," in *Proceedings of the ACL 2010 Conference Short Papers*, (Uppsala, Sweden), pp. 225--230, Association for Computational Linguistics, July 2010.

References

- P. Blunsom and T. Cohn, ``Inducing synchronous grammars with slice sampling," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, (Los Angeles, California), pp. 238--241, Association for Computational Linguistics, June 2010.
- A. Bouchard-Coˆte ́, S. Petrov, and D. Klein, ``Randomized pruning: Efficiently calculating expectations in large dynamic programs," in *Advances in Neural Information Processing Systems 22*.
- S. Kumar and W. Byrne, ``Minimum bayes-risk decoding for statistical machine translation," in *HLT-NAACL 2004: Main Proceedings* (D. M. Susan Dumais and S. Roukos, eds.), (Boston, Massachusetts, USA), pp. 169--176, Association for Computational Linguistics, May 2 - May 7 2004.
- R. Tromble, S. Kumar, F. Och, and W. Macherey, ``Lattice Minimum Bayes-Risk decoding for statistical machine translation," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, (Honolulu, Hawaii), pp. 620--629, Association for Computational Linguistics, October 2008.