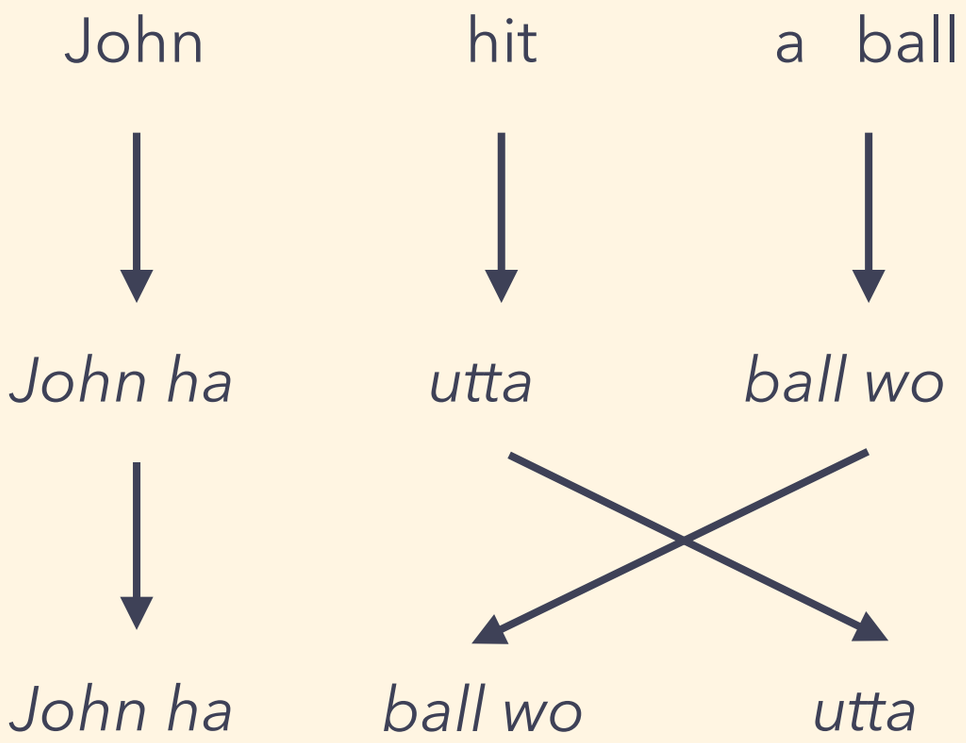
Rule-based Syntactic Preprocessing for Syntax-based Machine Translation

Nara Institute of Science and Technology, Japan
- Yuto Hatakoshi
- Graham Neubig
- Sakriani Sakti
- Tomoki Toda
- Satoshi Nakamura

25 October 2014 , Doha , Qatar

Phrase-based Machine Translation (PBMT)



Translation Model

Reordering Model

- Doesn’t incorporate syntactic information
- Difficulty estimating the probability of long distance reordering

Preprocessing using syntactic information

Lexical Processing for PBMT

Change the words in the sentence

Example: Reduce errors in verb conjugation and noun case agreement

[Avramidis and Koehn, 2008]

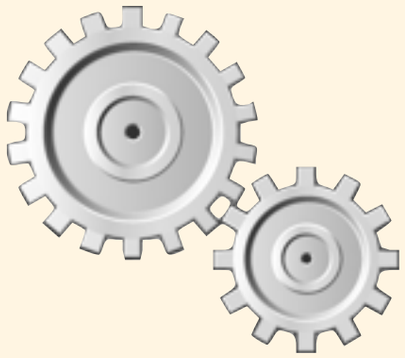
Annotation

[we: nominative] resolved the [issue: dative] of … or [relations: dative] with Serbia

Source language (Tagged)

Target language

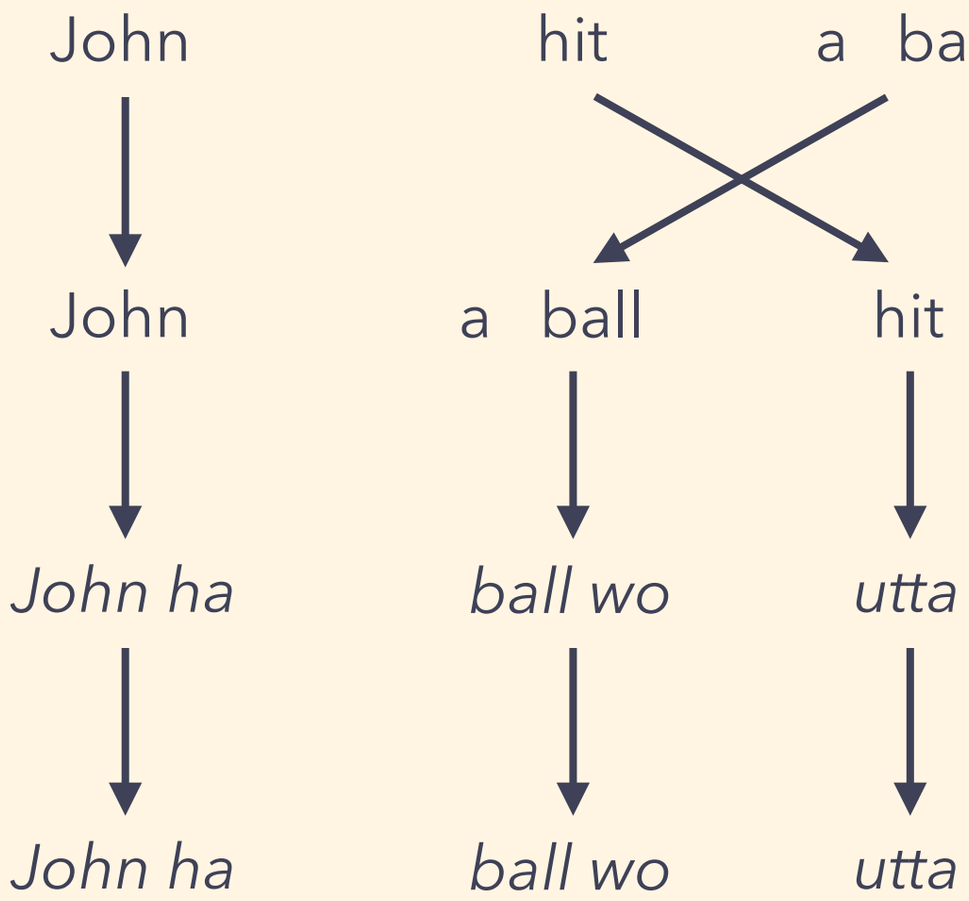
PBMT



Pre-ordering

[Xia and McCord, 2004]

Rearrange source sentence into target language word order



Preordering method

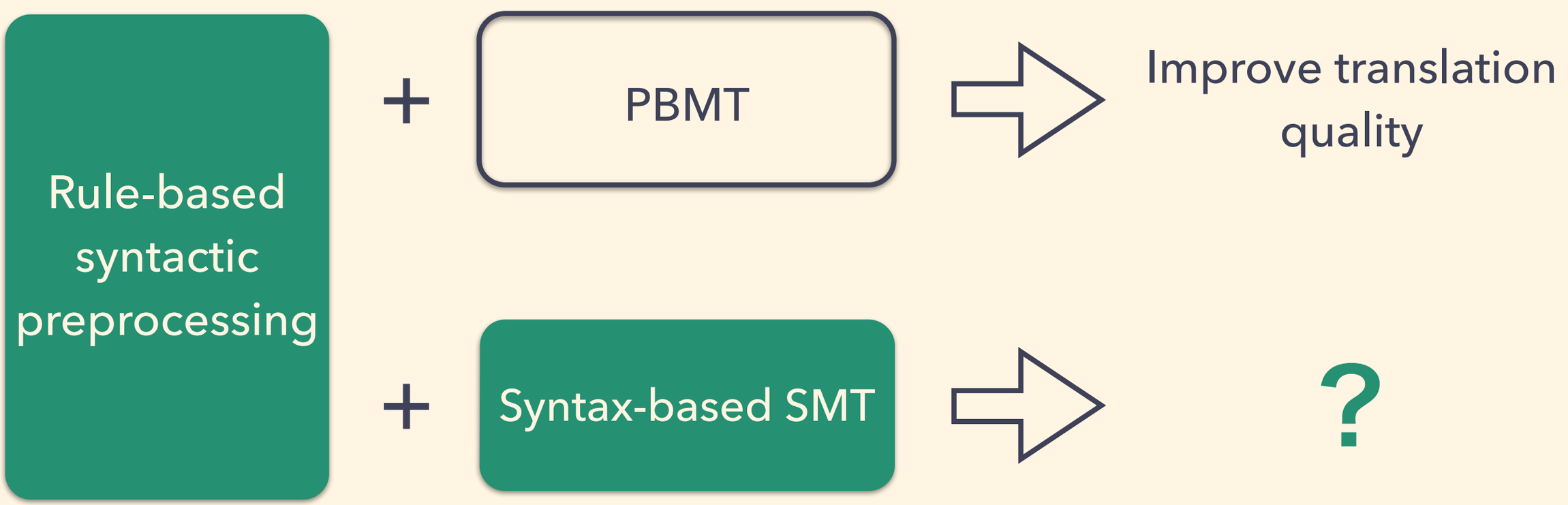
hand-crafted or trained rules

Translation Model

Reordering Model

Motivation

- Rule-based syntactic preprocessing is useful for PBMT
- Few attempts have been made for Syntax-based SMT
- Examine whether it also can contribute to Syntax-based SMT



Head Finalization:
A Syntactic Preprocessing Method for PBMT

Head Finalization

[Isozaki et al., 2010]

- Syntactic preprocessing method for **English to Japanese PBMT**
- Show significant improvements through 2 steps

1

Reordering

Convert English sentence into Japanese word order

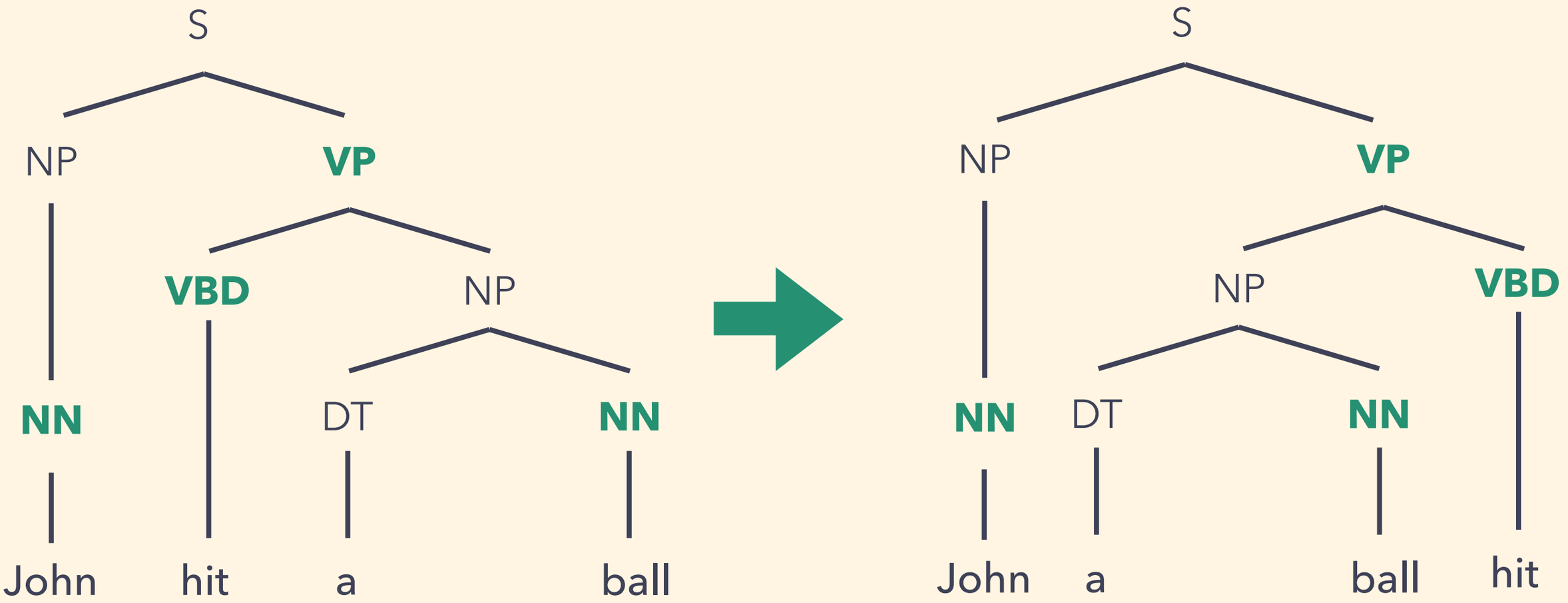
2

Lexical Processing

Generate more Japanese-like sentences

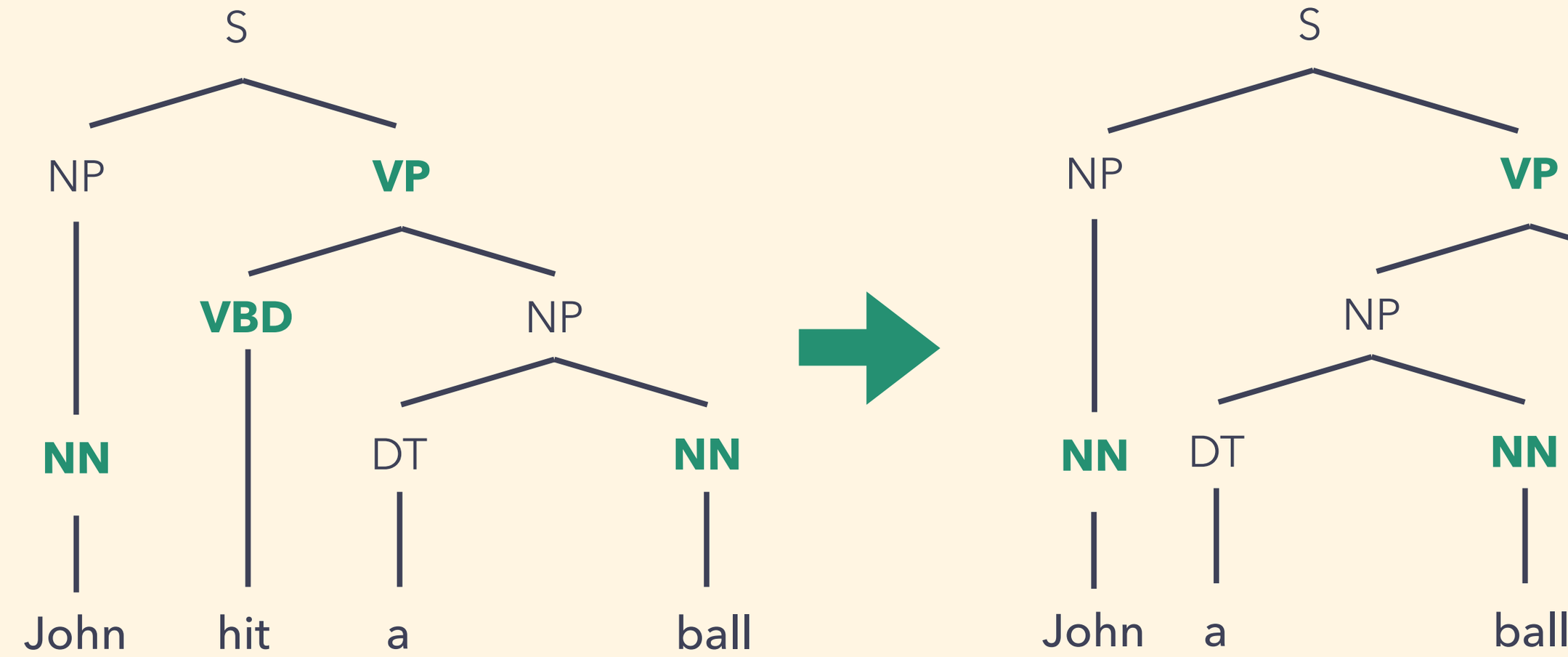
Reordering

1. Move head words to the end of the corresponding syntactic constituents



Reordering

1. Move head words to the end of the corresponding syntactic constituents



Japanese word order

Lexical Processing

2 Generate sentences closer to Japanese

Determinar elimination / Singularization

John ~~a~~ ball hit

Pseudo-particle insertion

John ***_wa_*** ball ***_wo_*** hit

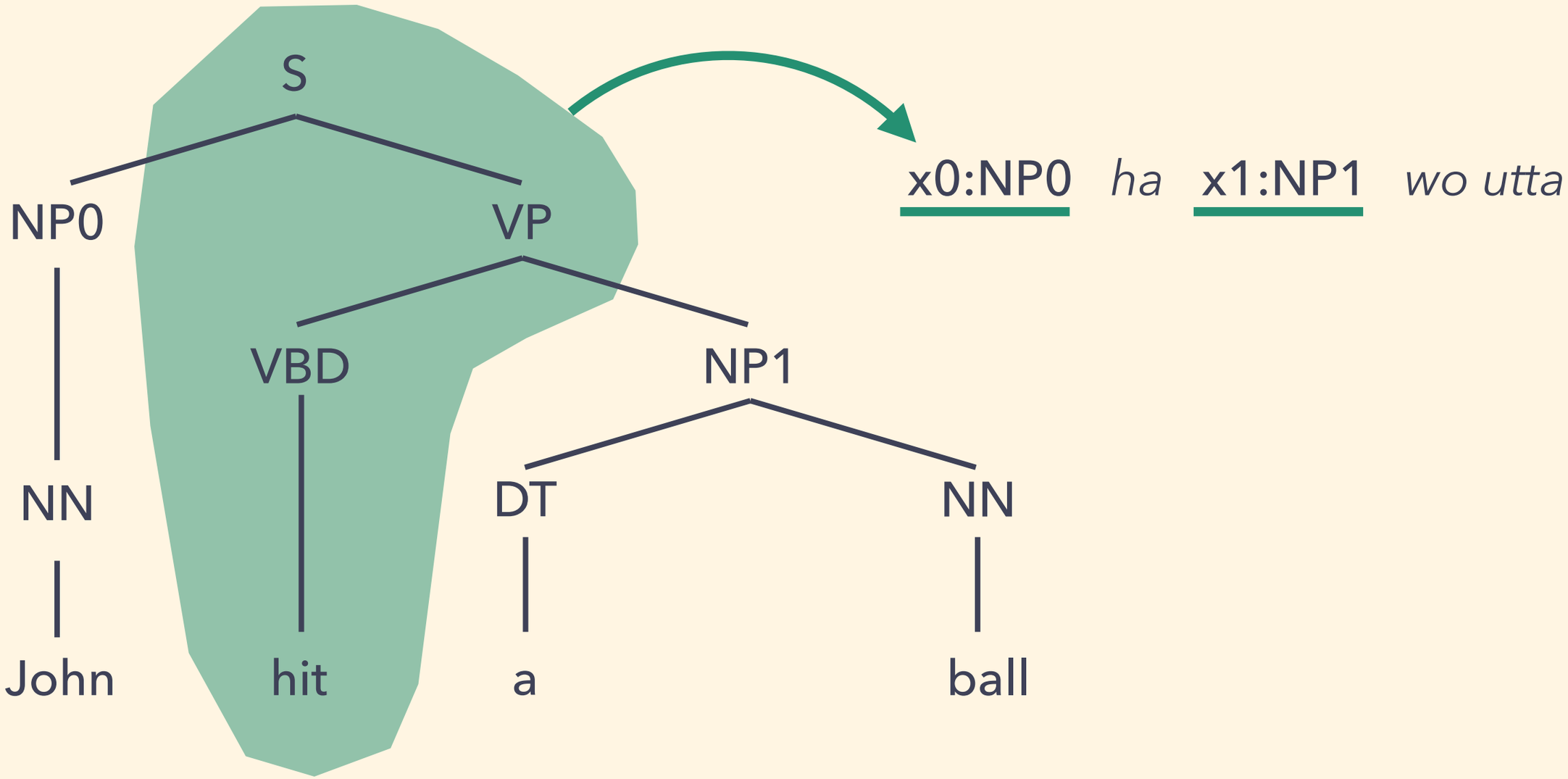
- _wa_ : Subject particle of the main verb
- _ga_ : Subject particle of other verb
- _wo_ : Object particle of any verb

Syntactic Preprocessing for T2S

Tree-to-String machine translation (T2S)

[Liu et al., 2006]

- Use parsing results of the source sentence
  - Possible to generate translations that are more accurate
  - Possible to handle long distance reordering



Potential Effect of Preprocessing on T2S

Reordering

- Improve word alignment
- Identify good translation patterns

Lexical Processing

- Improve translation quality of words

Proposed method

Apply three methods to **T2S**:

1

Reordering

Convert English sentence into Japanese word order

2

Lexical Processing

Determiner elimination / Singularization / Particle insertion

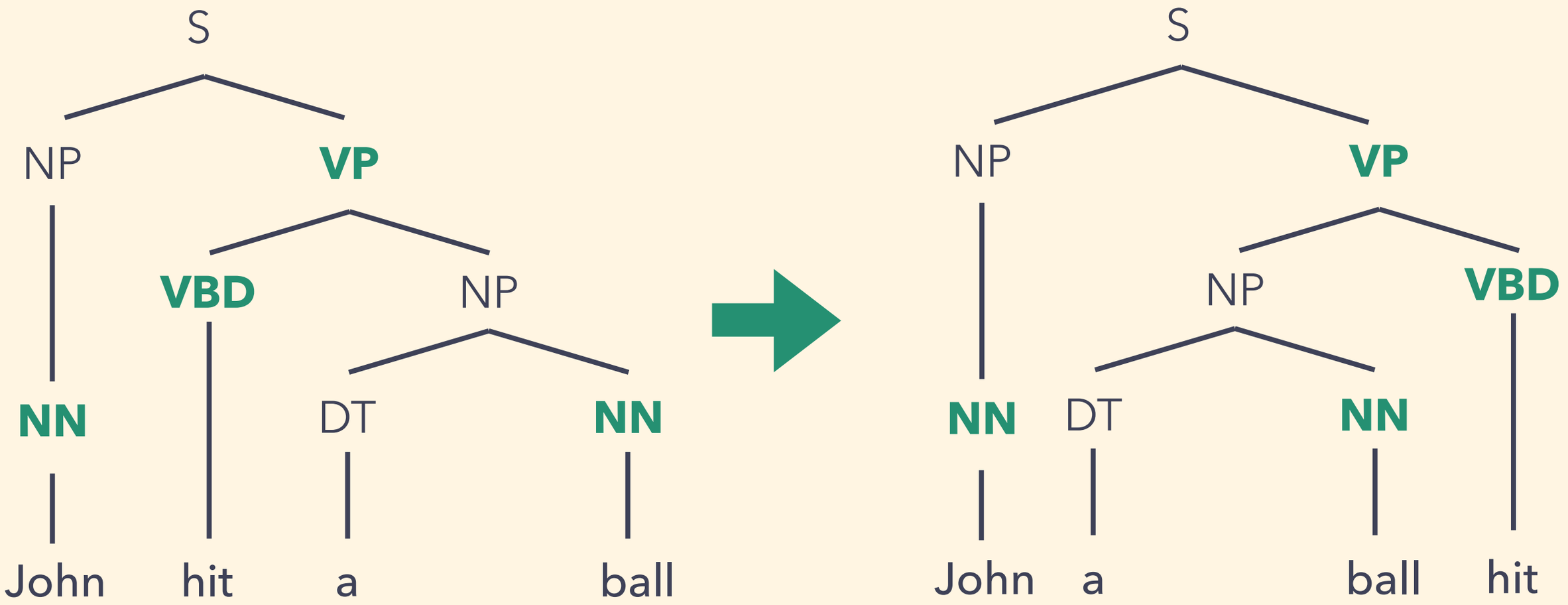
3

HF-feature

Apply reordering information to T2S as soft constraints

Reordering for T2S

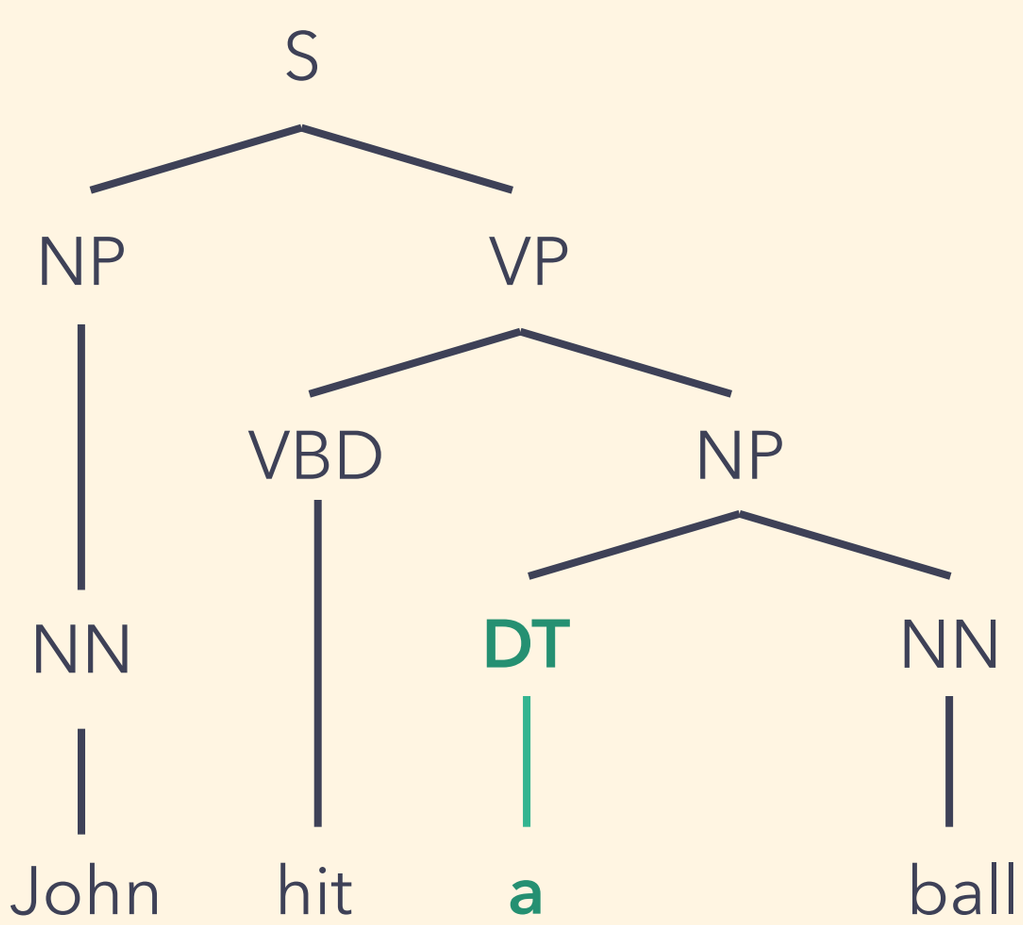
- Convert English sentence into Japanese word order
- Output the reordered tree

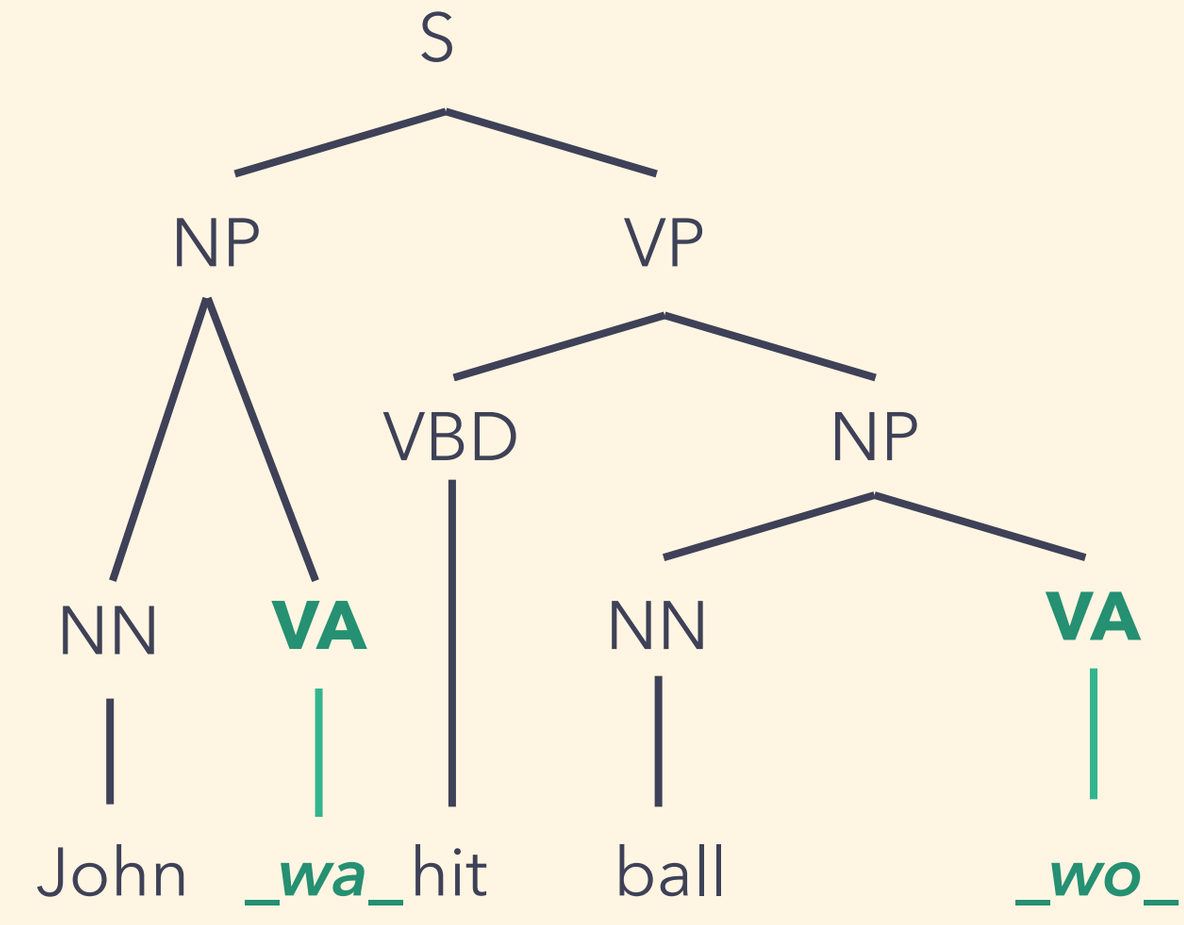


Improve word alignment

Lexical Processing for T2S

- Determinar elimination / Singularization / Particle insertion
- Transform not strings, but trees





Improve translation performance of word

Reordering Information as Soft Constraints

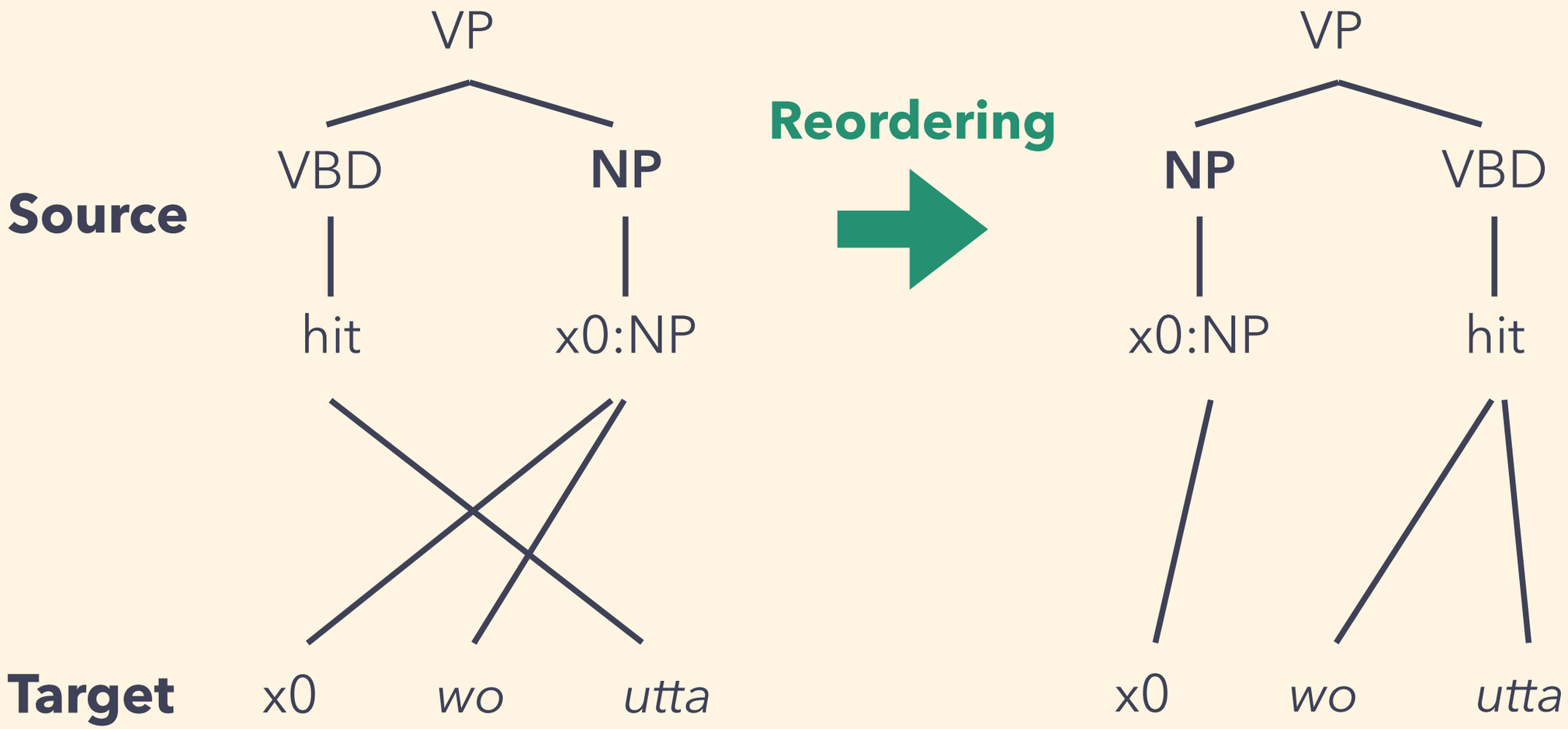
- Some translation patterns do not obey head final order due to bad alignment
- Sometimes head final order is not applicable in Japanese grammar

- Log-linear model

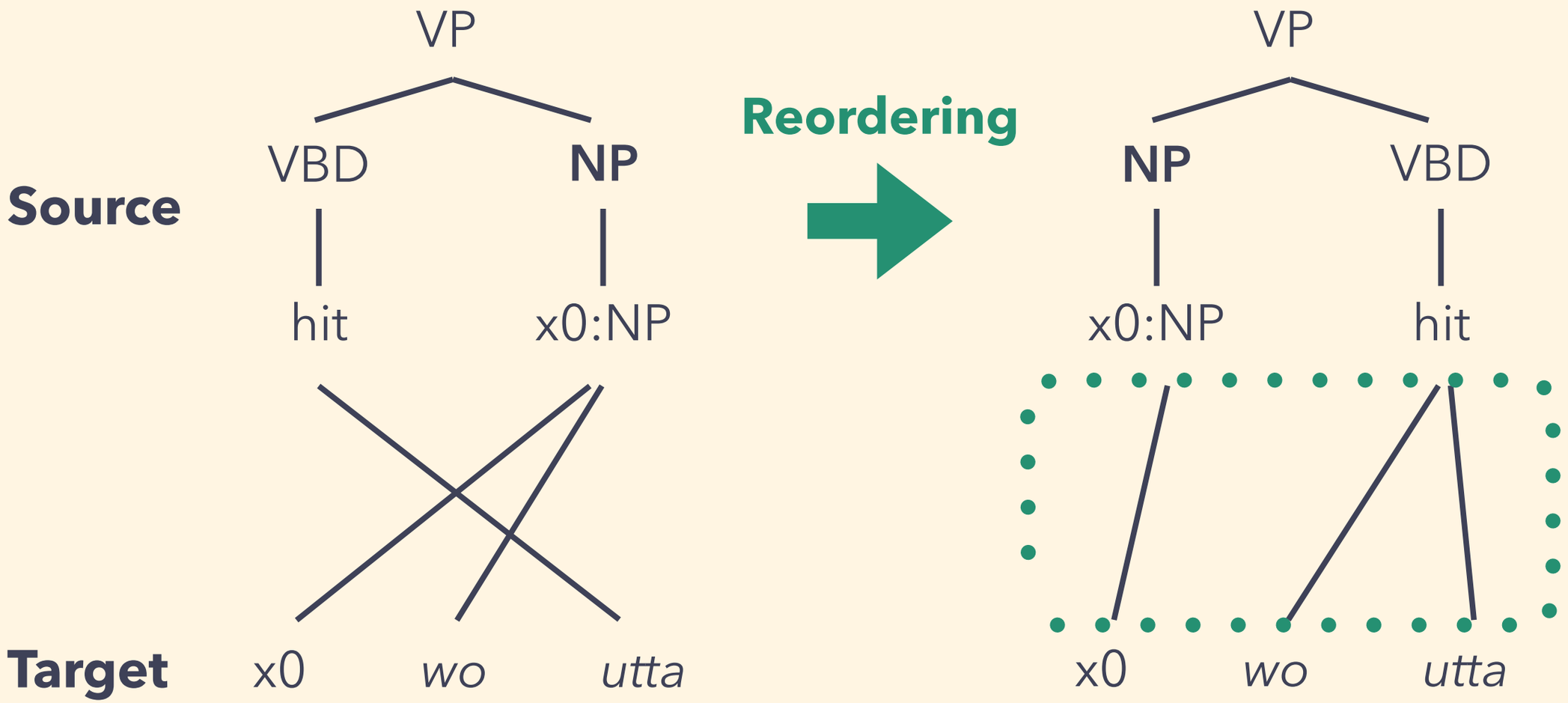
$$\hat{e} = \underset{e}{\operatorname{argmax}}\, \boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{h}(\boldsymbol{f}, \boldsymbol{e})$$

$\boldsymbol{f}$	source sentence
$\boldsymbol{e}$	target sentence
$\boldsymbol{h}(\cdot)$	feature function

Procedure of HF-feature addition



Procedure of HF-feature addition



Non-crossing $h_{\mathrm{HF}}(\boldsymbol{f}, \boldsymbol{e}) = 1$

Crossing $h_{\mathrm{HF}}(\boldsymbol{f}, \boldsymbol{e}) = 0$

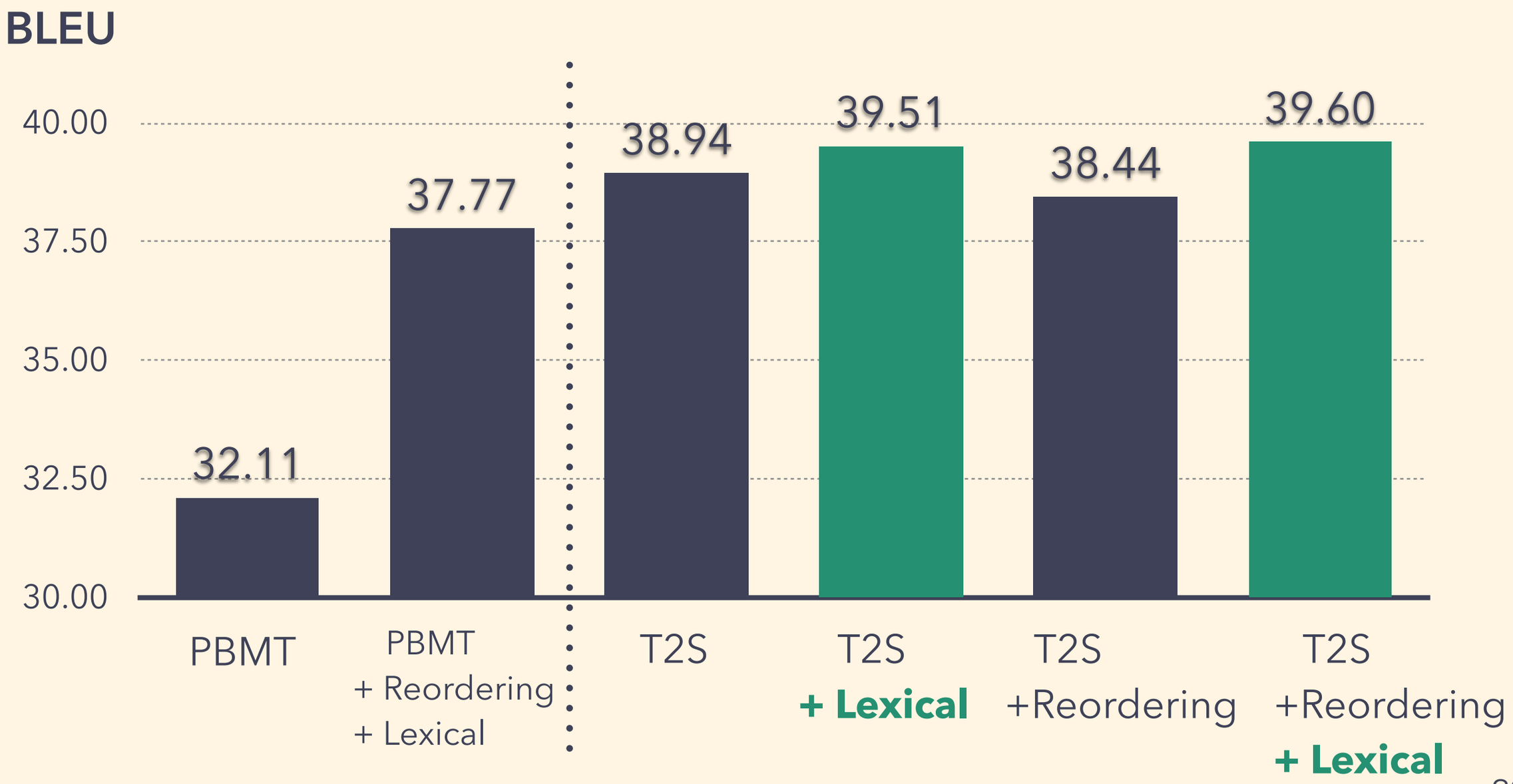
Experiment and Result

Experimental Environment

- Translation Task
  - English → Japanese
  - NTCIR-7 (train: 3.08M, dev: 0.82k, test: 1.38k sentences)
- Translation Method
  - PBMT (default settings of moses)
  - T2S (default settings of travatar)
- Evaluation
  - BLEU, RIBES

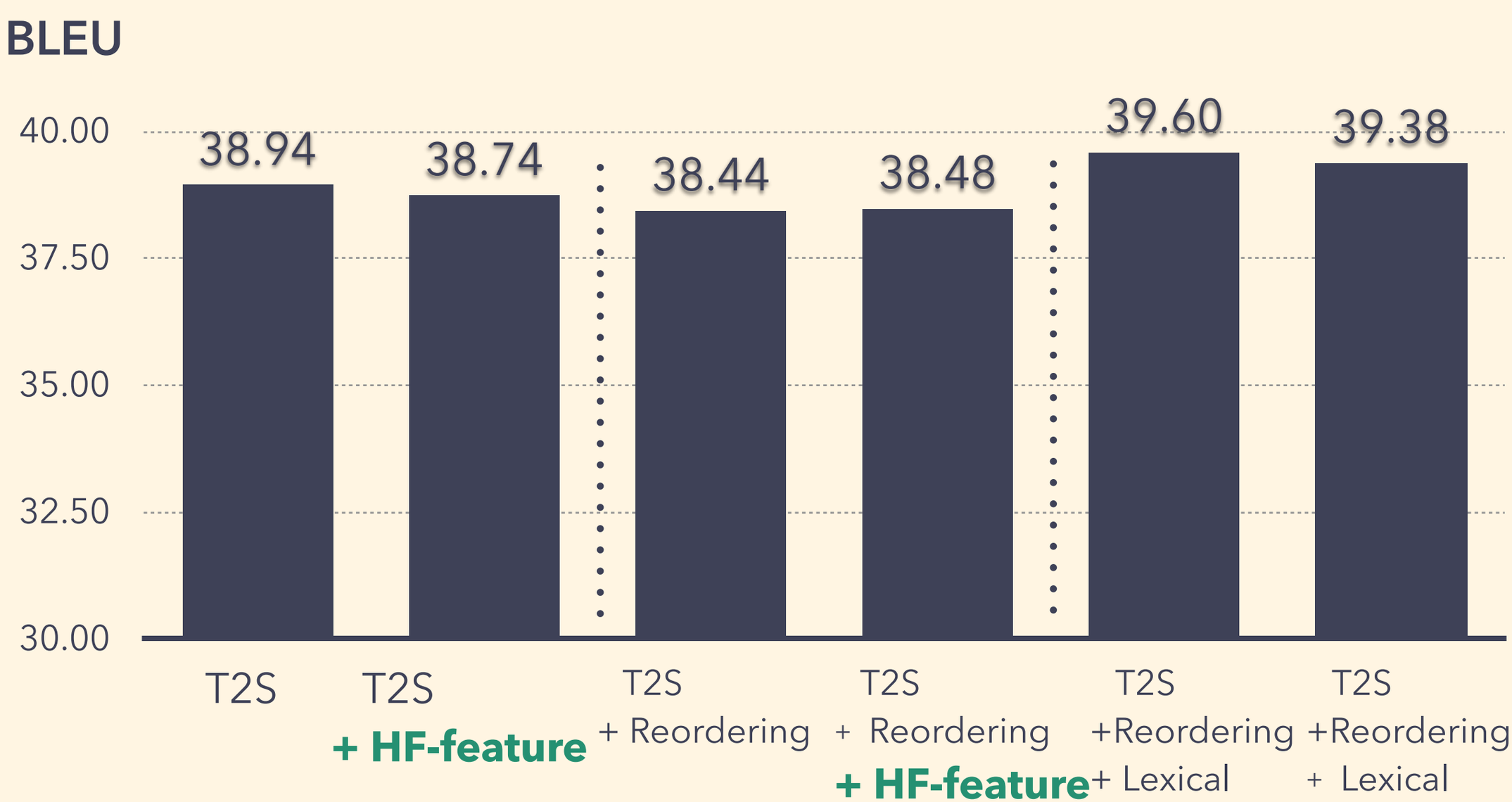
Translation quality

- Translation quality is improved by Lexical Processing
- Reordering is not effective for T2S



Translation quality

- Translation quality is not improved by HF-feature
  - Reordering quality achieved by T2S was already high



Conclusion and Future Work

- Applied rule-based syntactic preprocessing designed for PBMT to T2S

Reordering +

Lexical Processing +

Reordering feature ⇨

- Examine other language pairs / Apply preprocessing to F2S

Improvement by Lexical Processing

◉ - Lexical Processing

図 に 示 す よう に 、 電気 絶縁 性 の ハウジング ９７に 一 列 に 並 ぶ 複数 の 雄型 コンタクト ９８ と から 構成 され て い る 。

s (x0:np vp (pn (, (“,”)) vp (vx (vbz (“comprises”)) x1:np)))

→ x0 x1 “と” “から” “構成” “さ” “れ” “て” “い”

◉ + Lexical Processing

図 に 示 す よう に 、 電気 絶縁 性 の ハウジング ９７に 一 列 に 並 ぶ 複数 の 雄型 コンタクト ９８ を 有 し て 構成さ れ る 。

np (np (x0:nx) va (_va2 “_va2”)) → x0 “を”

Optimized weight of HF-feature

- HF-feature led to confusion in MERT optimization
- There is no consistent pattern of learning weights

HF-feature	Reordering	Lexical	Weight of HF-feature
–	–	–	-0.00707078
–	–	+	0.00524676
–	+	–	0.156724
–	+	+	-0.121326

Translation Quality

			PBMT		T2S	
HF-feature	Reordering	Lexical	BLEU	RIBES	BLEU	RIBES
–	–	–	32.11	69.06	38.94	78.48
–	–	+	33.16	70.19	**39.51**	**79.47**
–	+	–	37.62	77.56	38.44	78.48
–	+	+	37.77	77.71	**39.60**	**79.26**
+	–	–	–	–	38.74	78.33
+	–	+	–	–	**39.29**	**79.23**
+	+	–	–	–	38.48	78.44
+	+	+	–	–	**39.38**	**79.21**