Chunk-based Statistical Translation

Taro Watanabe†, Eiichiro Sumita and Hiroshi G. Okuno

taro.watanabe@atr.co.jp

ATR Spoken Language Translation Research Laboratories

Contents

Statistical Machine Translation

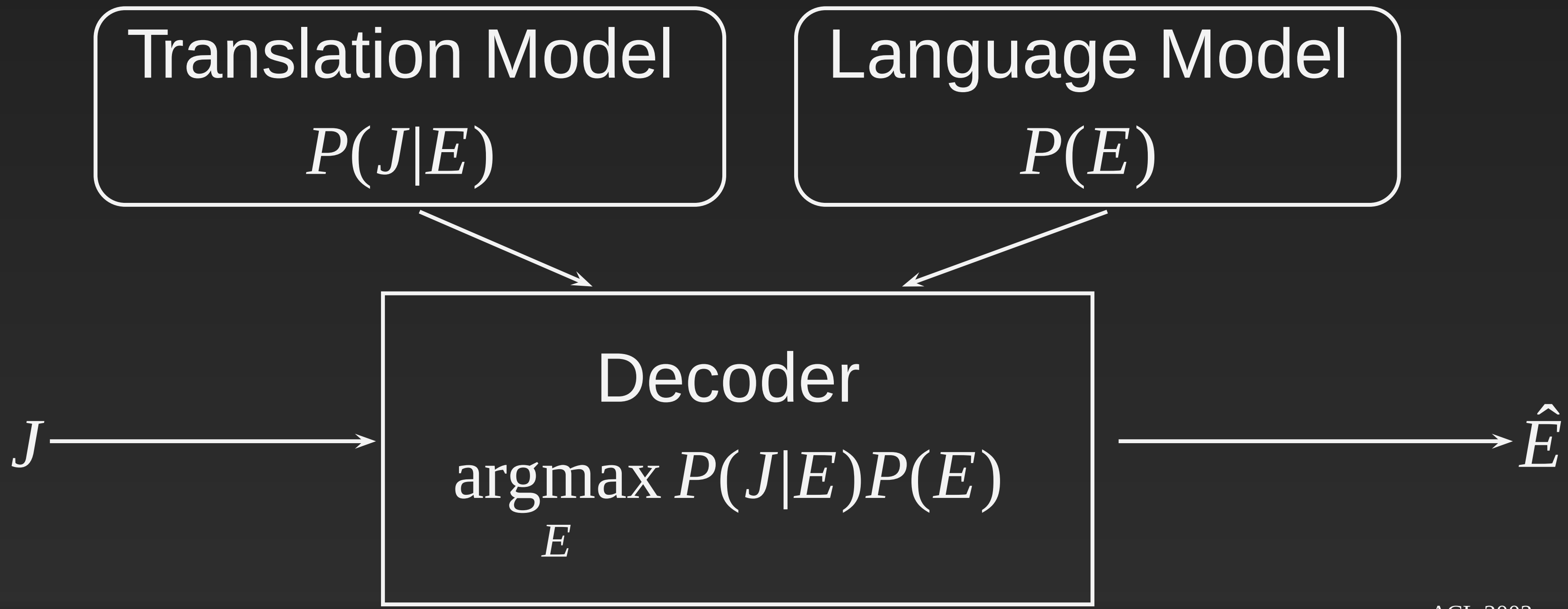
- Translation from J into E

$$
\begin{aligned}
\hat{E} &= \operatorname*{argmax}_{E} P(E|J) \\
&= \operatorname*{argmax}_{E} P(E)P(J|E)
\end{aligned}
$$

Statistical Machine Translation

- Translation from J into E

$$\hat{E} = \underset{E}{\operatorname{argmax}} P(E|J)$$
$$= \underset{E}{\operatorname{argmax}} P(E)P(J|E)$$



Word Alignment Based Statistical Translation

$$P(J|E) = \sum_{A} P(J, A|E)$$

$E =$ NULL$_0$ show$_1$ me$_2$ the$_3$ one$_4$ in$_5$ the$_6$ window$_7$

$J =$ uindo$_1$ no$_2$ shinamono$_3$ o$_4$ mise$_5$ tekudasai$_6$

$A = ($ 7 0 4 0 1 1 $)$

Word Alignment Based Statistical Translation

$$P(J|E) = \sum_{A} P(J, A|E)$$

$E =$ NULL$_0$ show$_1$ me$_2$ the$_3$ one$_4$ in$_5$ the$_6$ window$_7$

$J =$ uindo$_1$ no$_2$ shinamono$_3$ o$_4$ mise$_5$ tekudasai$_6$

$A = ($ 7 0 4 0 1 1 $)$

- Generative Process of $P(J, A|E)$

An Example — IBM Model 4

Fertility

$n(2|E_1)$

$n(0|E_2)$

$n(0|E_3)$

...

NULL

$\binom{4}{2} p_0^{4-2} p_1^2$

Lexicon

$t(J_5|E_1)$

$t(J_6|E_1)$

$t(J_3|E_4)$

...

Distortion

$d_1(1 - \lceil \frac{3}{1} \rceil | E_4, J_1)$

$d_1(3 - \lceil \frac{5+6}{2} \rceil | E_1, J_3)$

$d_1(5 - \lceil \frac{2+4}{2} \rceil | \text{NULL}, J_5)$

$d_{>1}(6 - 5 | J_6)$

An Example — IBM Model 4

Fertility

$n(2|E_1)$

$n(0|E_2)$

$n(0|E_3)$

...

NULL

$\binom{4}{2} p_0^{4-2} p_1^2$

Lexicon

$t(J_5|E_1)$

$t(J_6|E_1)$

$t(J_3|E_4)$

...

Distortion

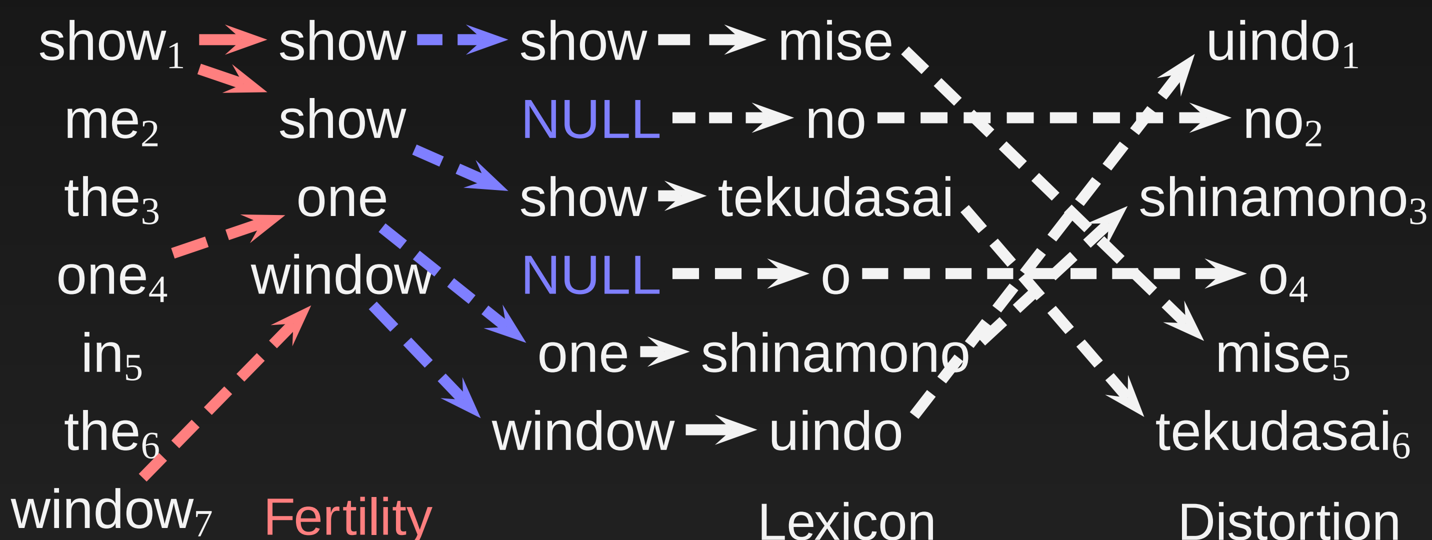
$d_1(1 - \lceil \frac{3}{1} \rceil | E_4, J_1)$

$d_1(3 - \lceil \frac{5+6}{2} \rceil | E_1, J_3)$

$d_1(5 - \lceil \frac{2+4}{2} \rceil | \text{NULL}, J_5)$

$d_{>1}(6 - 5 | J_6)$

An Example — IBM Model 4



Fertility

$n(2|E_1)$

$n(0|E_2)$

$n(0|E_3)$

...

NULL

$\binom{4}{2} p_0^{4-2} p_1^2$

Lexicon

$t(J_5|E_1)$

$t(J_6|E_1)$

$t(J_3|E_4)$

...

Distortion

$d_1(1 - \lceil \frac{3}{1} \rceil | E_4, J_1)$

$d_1(3 - \lceil \frac{5+6}{2} \rceil | E_1, J_3)$

$d_1(5 - \lceil \frac{2+4}{2} \rceil | \mathrm{NULL}, J_5)$

$d_{>1}(6 - 5 | J_6)$

An Example — IBM Model 4

Fertility

$n(2|E_1)$

$n(0|E_2)$

$n(0|E_3)$

...

NULL

$\binom{4}{2} p_0^{4-2} p_1^2$

Lexicon

$t(J_5|E_1)$

$t(J_6|E_1)$

$t(J_3|E_4)$

...

Distortion

$d_1(1 - \lceil \frac{3}{1} \rceil | E_4, J_1)$

$d_1(3 - \lceil \frac{5+6}{2} \rceil | E_1, J_3)$

$d_1(5 - \lceil \frac{2+4}{2} \rceil | \text{NULL}, J_5)$

$d_{>1}(6 - 5 | J_6)$

An Example — IBM Model 4

Fertility

$n(2|E_1)$

$n(0|E_2)$

$n(0|E_3)$

...

NULL

$\binom{4}{2} p_0^{4-2} p_1^2$

Lexicon

$t(J_5|E_1)$

$t(J_6|E_1)$

$t(J_3|E_4)$

...

Distortion

$d_1(1 - \lceil \frac{3}{1} \rceil | E_4, J_1)$

$d_1(3 - \lceil \frac{5+6}{2} \rceil | E_1, J_3)$

$d_1(5 - \lceil \frac{2+4}{2} \rceil | \text{NULL}, J_5)$

$d_{>1}(6 - 5 | J_6)$

Problems

- Strategy: Generate a set of words from each source word and reorder them.

Problems

- Strategy: Generate a set of words from each source word and reorder them.
- Insertion/Deletion Modeling
  - Fertility Model to select deletion
  - A binomial distribution to determine insertion

Problems

- Strategy: Generate a set of words from each source word and reorder them.
- Insertion/Deletion Modeling
  - Fertility Model to select deletion
  - A binomial distribution to determine insertion
- Local Alignment Modeling
  - Collection of Local Reordering $\longrightarrow$ Global Reodering
  - Long distance word alignment

Chunk-based Statistical Translation

$$P(J|E) = \sum_{\mathcal{J}} \sum_{\mathcal{E}} P(J, \mathcal{J}, \mathcal{E}|E)$$

$\mathcal{J}$, $\mathcal{E}$: sequences of chunks ($|\mathcal{J}| = |\mathcal{E}|$)

Chunk-based Statistical Translation

$$P(J|E) = \sum_{\mathcal{J}} \sum_{\mathcal{E}} P(J, \mathcal{J}, \mathcal{E}|E)$$

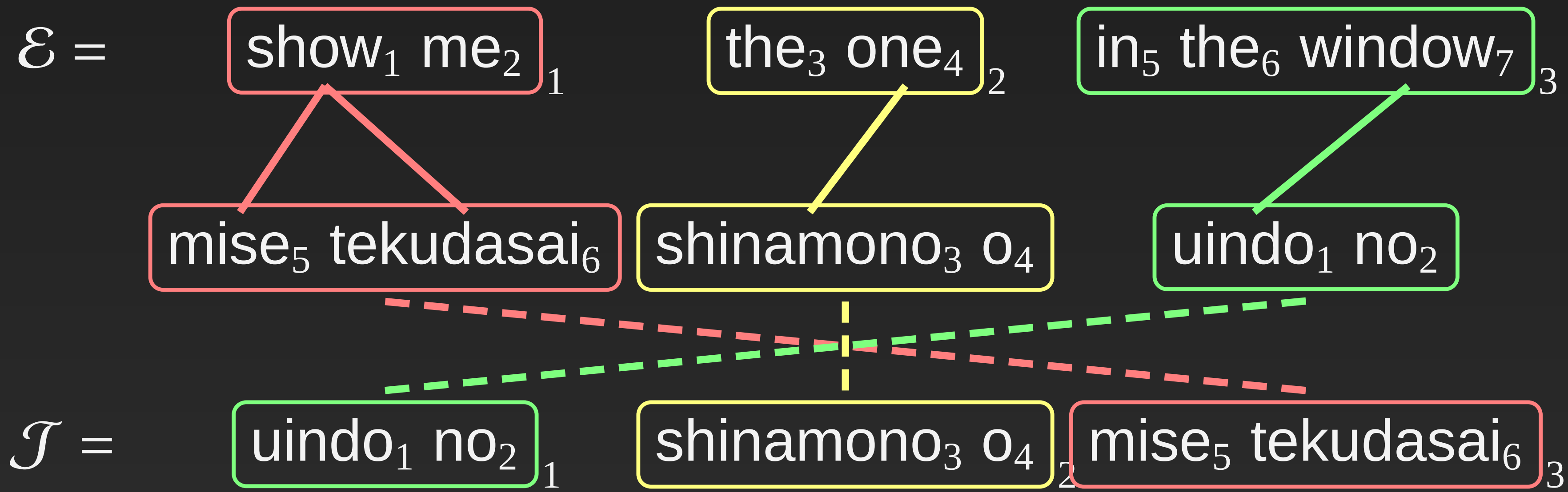
$$P(J, \mathcal{J}, \mathcal{E}|E) = \sum_{A} \sum_{\mathcal{A}} P(J, \mathcal{J}, A, \mathcal{A}, \mathcal{E}|E)$$

$\mathcal{J}$, $\mathcal{E}$: sequences of chunks ($|\mathcal{J}| = |\mathcal{E}|$)
A: chunk alignment
$\mathcal{A}$: word alignment

Chunk-based Statistical Translation

$$P(J|E) = \sum_{\mathcal{J}} \sum_{\mathcal{E}} P(J, \mathcal{J}, \mathcal{E}|E)$$

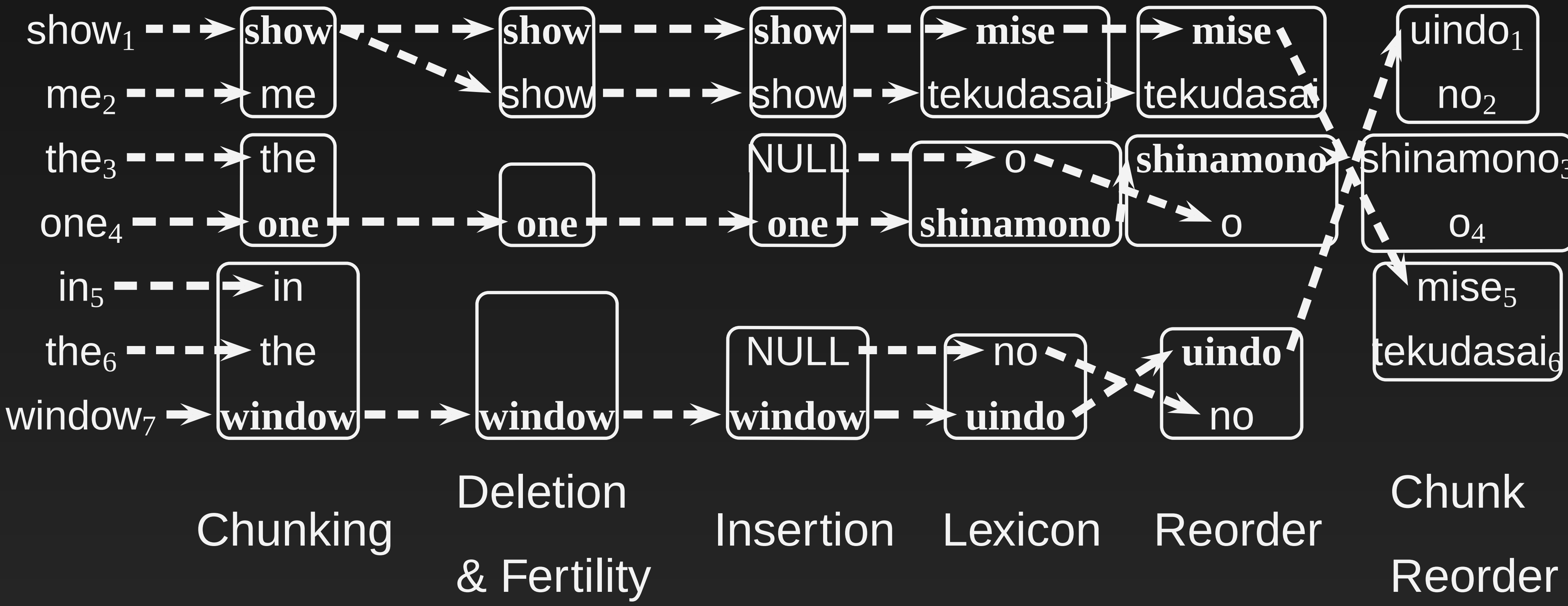
$$P(J, \mathcal{J}, \mathcal{E}|E) = \sum_{A} \sum_{\mathcal{A}} P(J, \mathcal{J}, A, \mathcal{A}, \mathcal{E}|E)$$



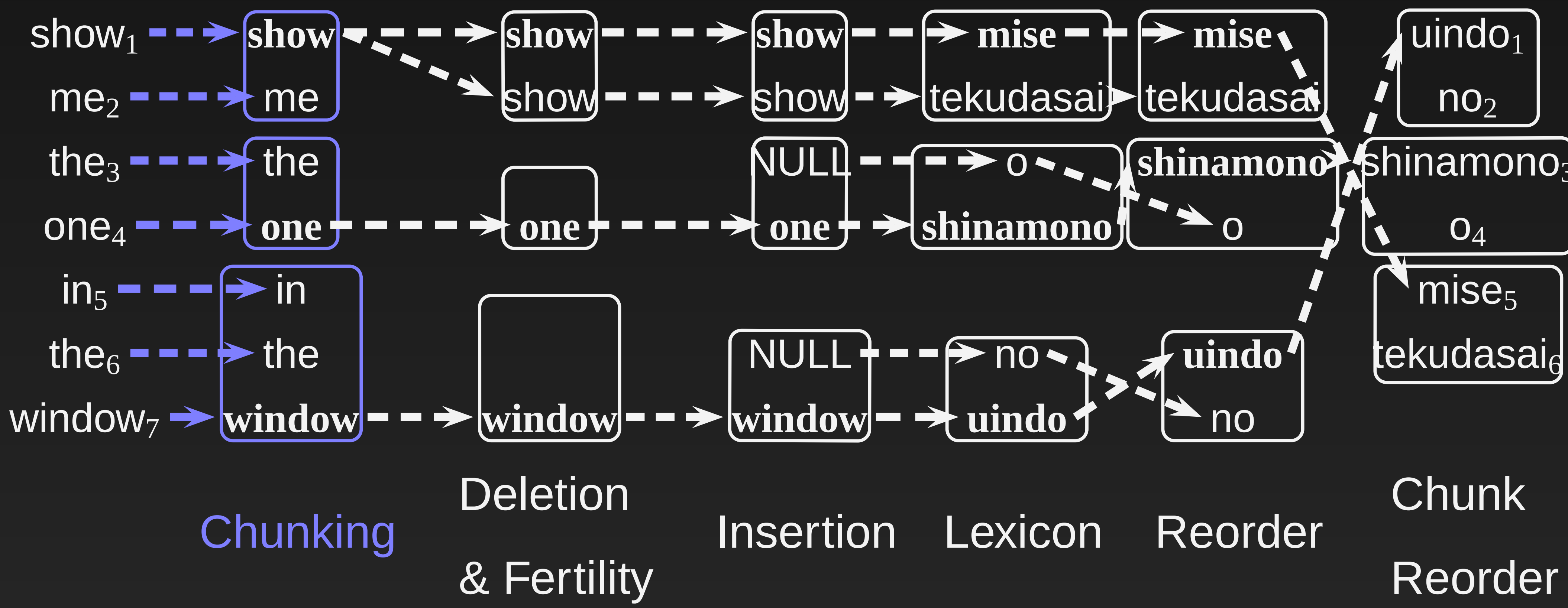
$A = (\quad 3 \quad 2 \quad 1 \quad)$

$\mathcal{A} = (\quad [7, 0] \quad [4, 0] \quad [1, 1] \quad)$

Model Structure

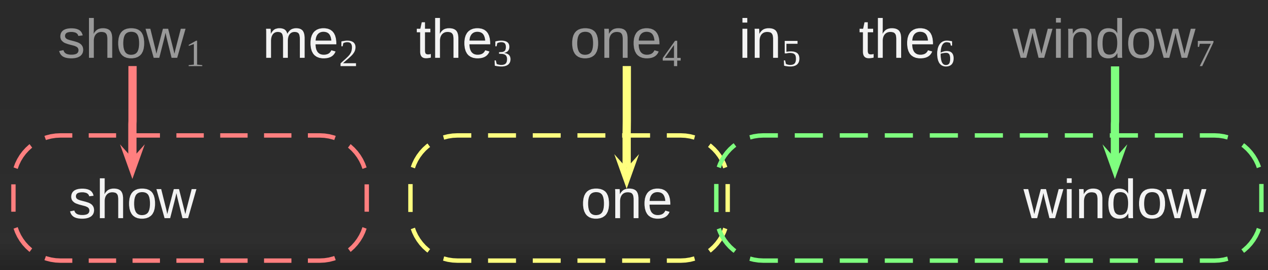


Model Structure

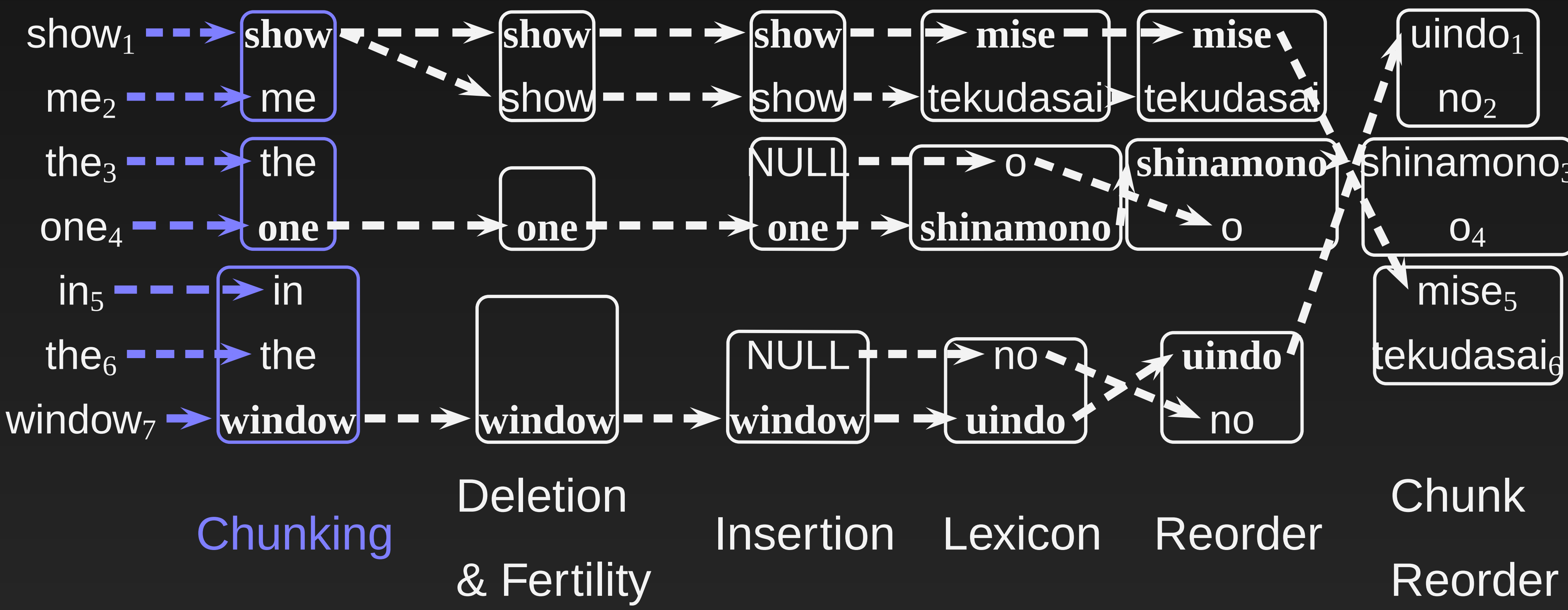


Chunking: Choose Chunk Size — $\prod_i \epsilon(\varphi_i | E_i)$

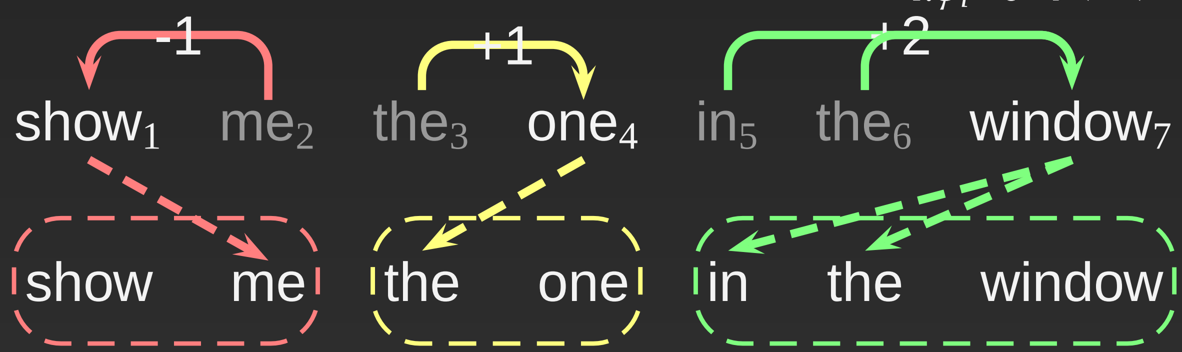
φ_i = chunk size and if $\varphi_i > 0$ then, E_i is a head word



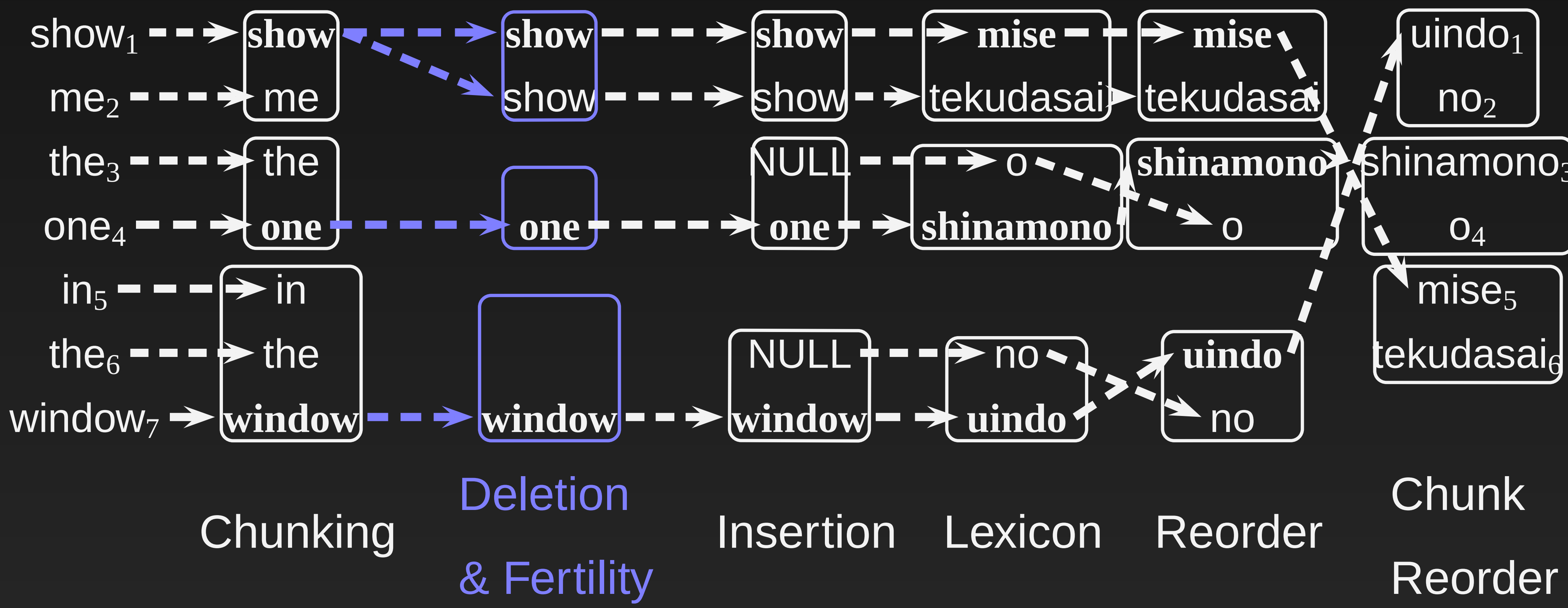
Model Structure



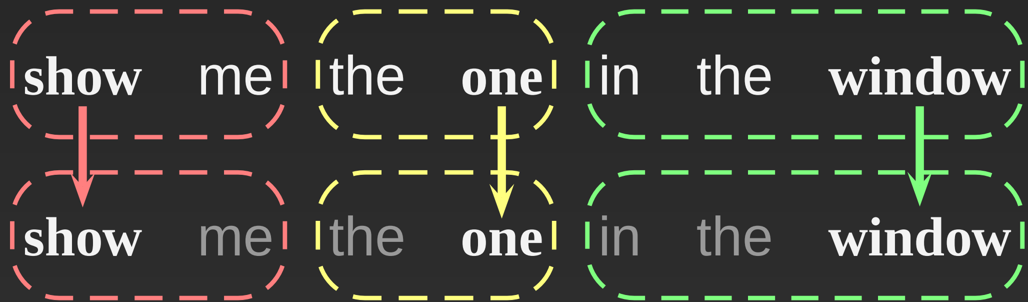
Chunking: Associate Non-Head Words — $\prod_{i:\varphi_i=0} \eta(c(E_{h_i})|h_i - i, c(E_i))$



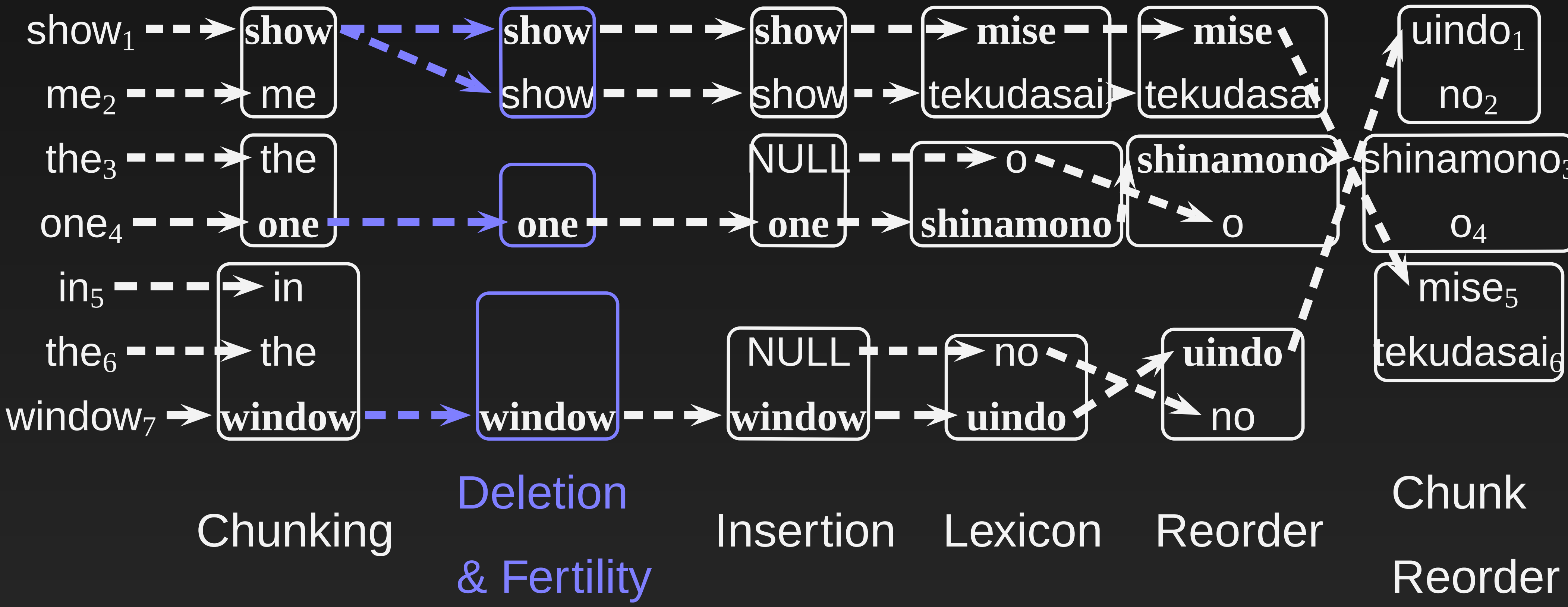
Model Structure



Deletion — $\prod_{i:\varphi_i=0} \delta(d_i|c(E_i), c(E_{h_i}))$

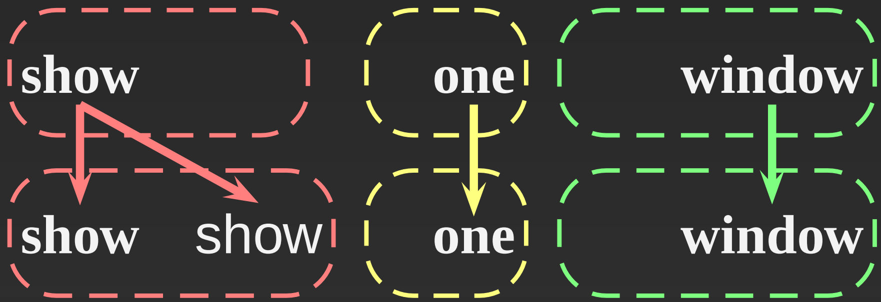


Model Structure

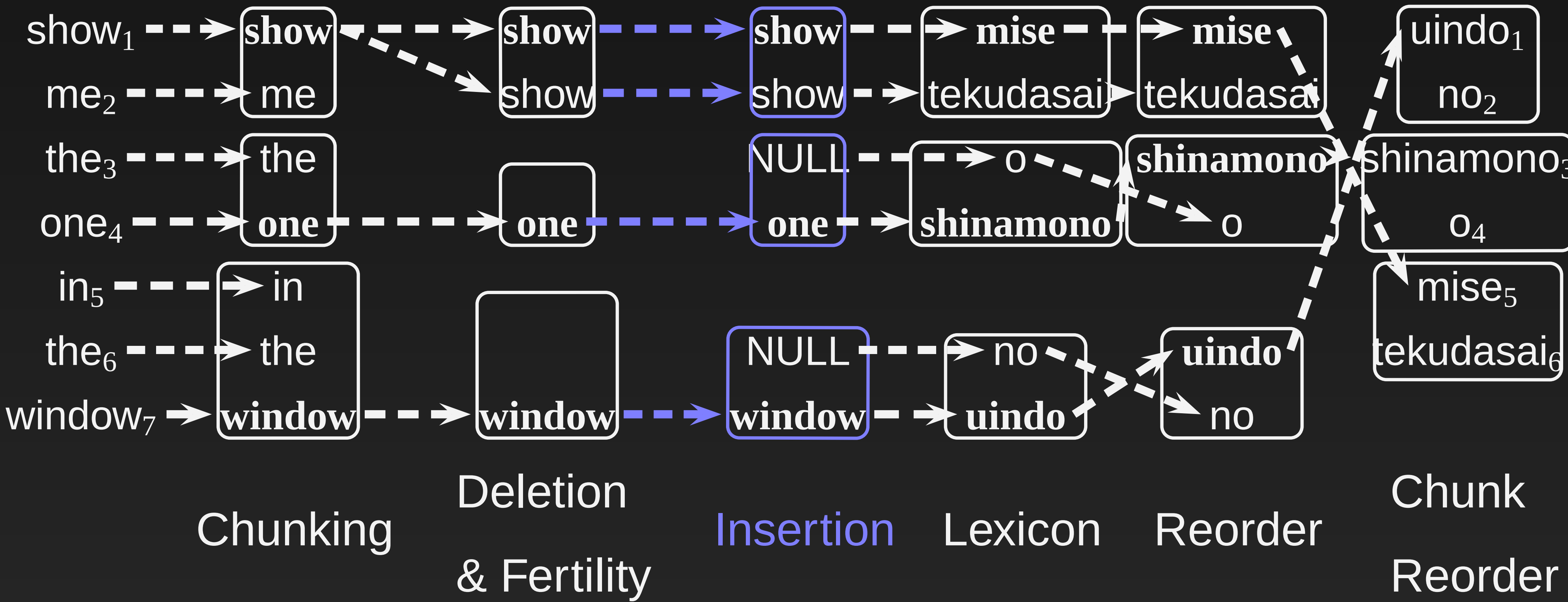


Fertility — $\prod_{i:\varphi_i>0} \nu(\phi_i|E_i)/\phi_i$

ϕ_i = # of words

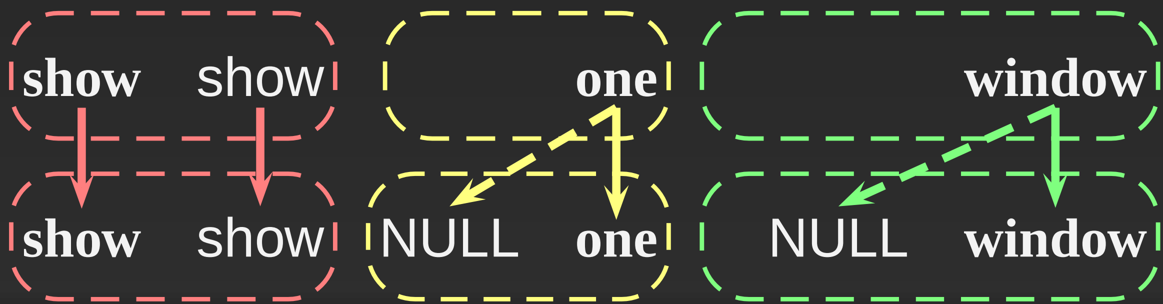


Model Structure

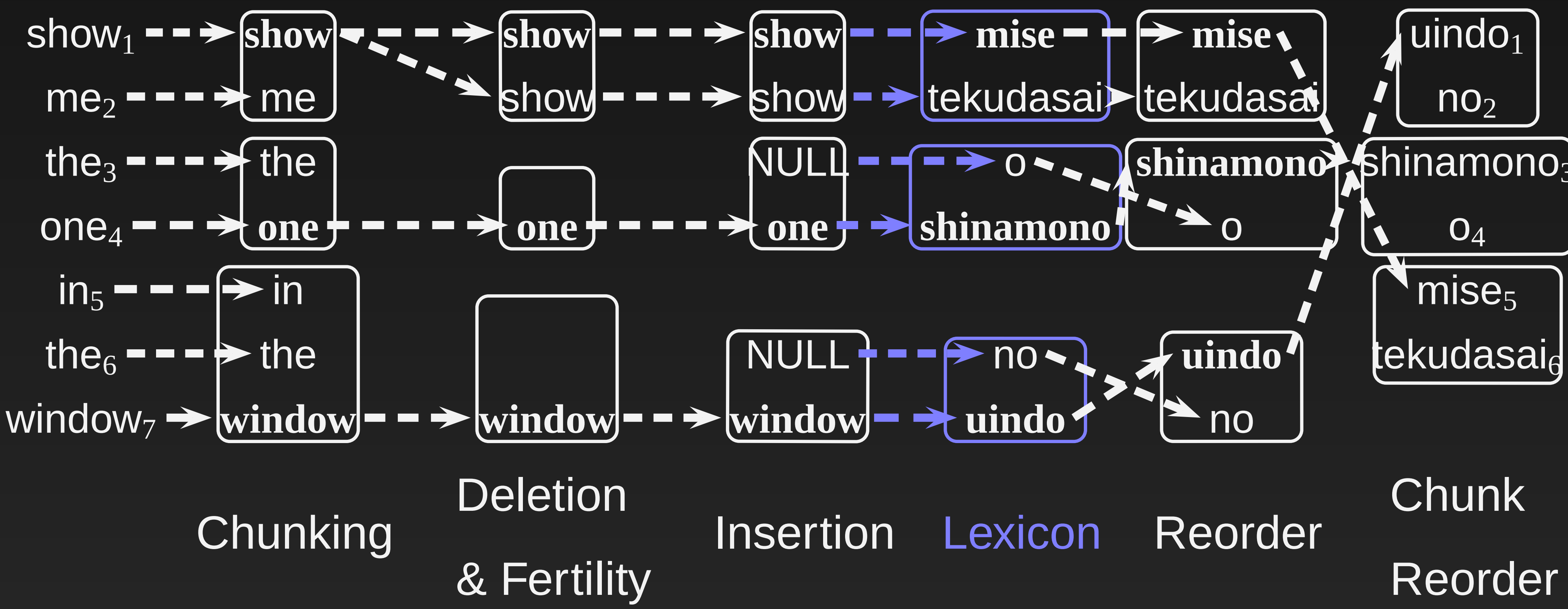


Insertion — $\prod_{i:\varphi_i>0} \iota(\phi'_i|c(E_i))$

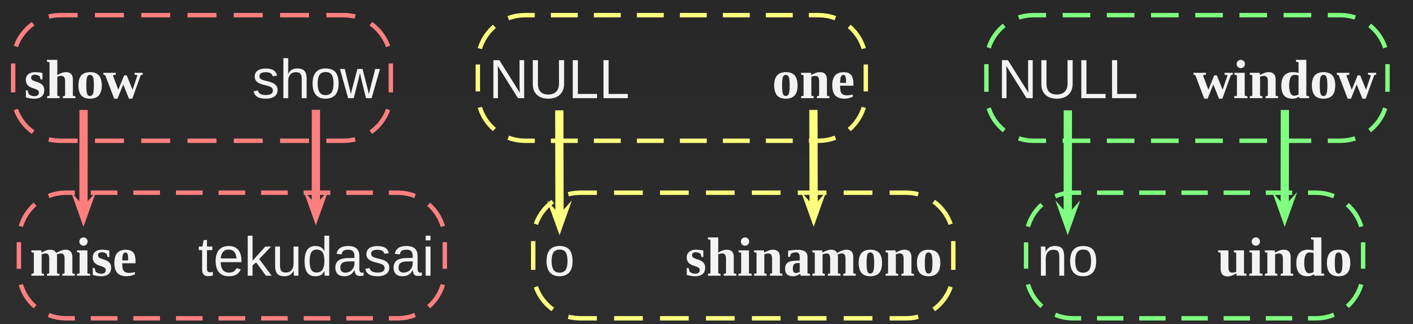
ϕ'_i = # of NULL words



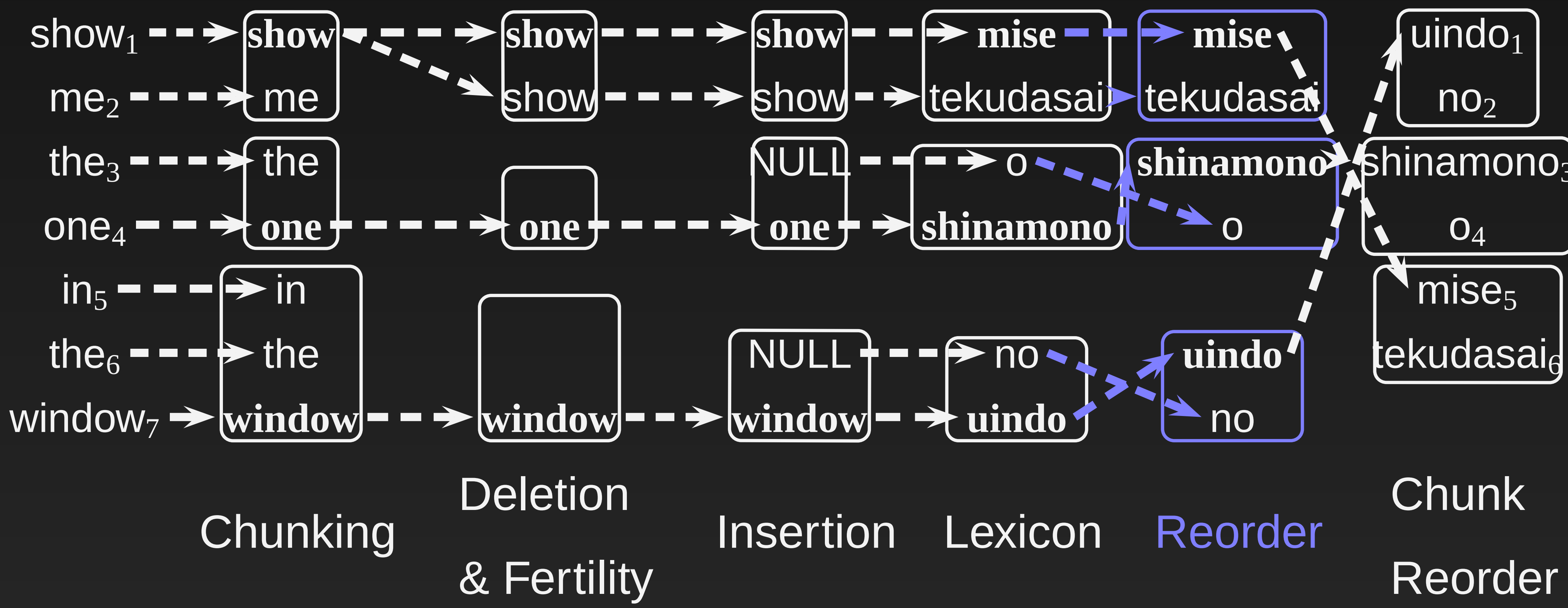
Model Structure



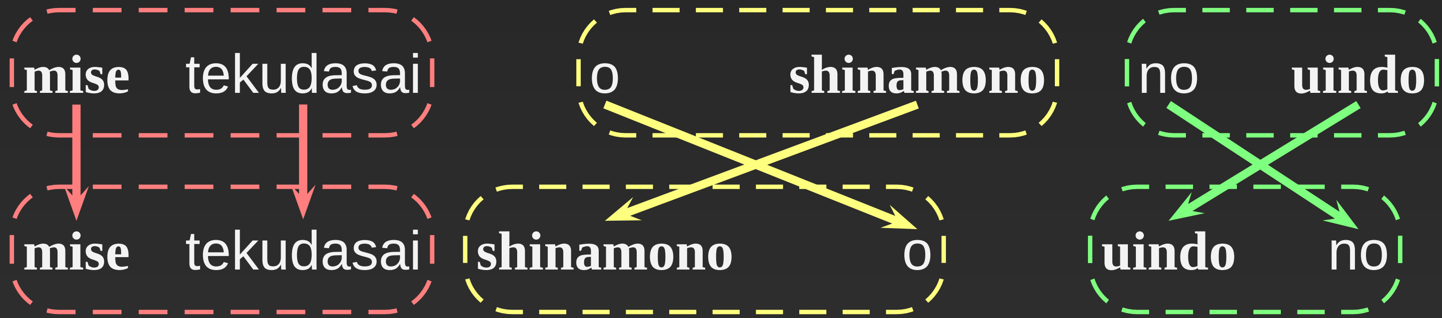
Lexical Transfer — $\prod_j \prod_k \tau(\mathcal{J}_{j,k}|E_{\mathcal{A}_{j,k}})$



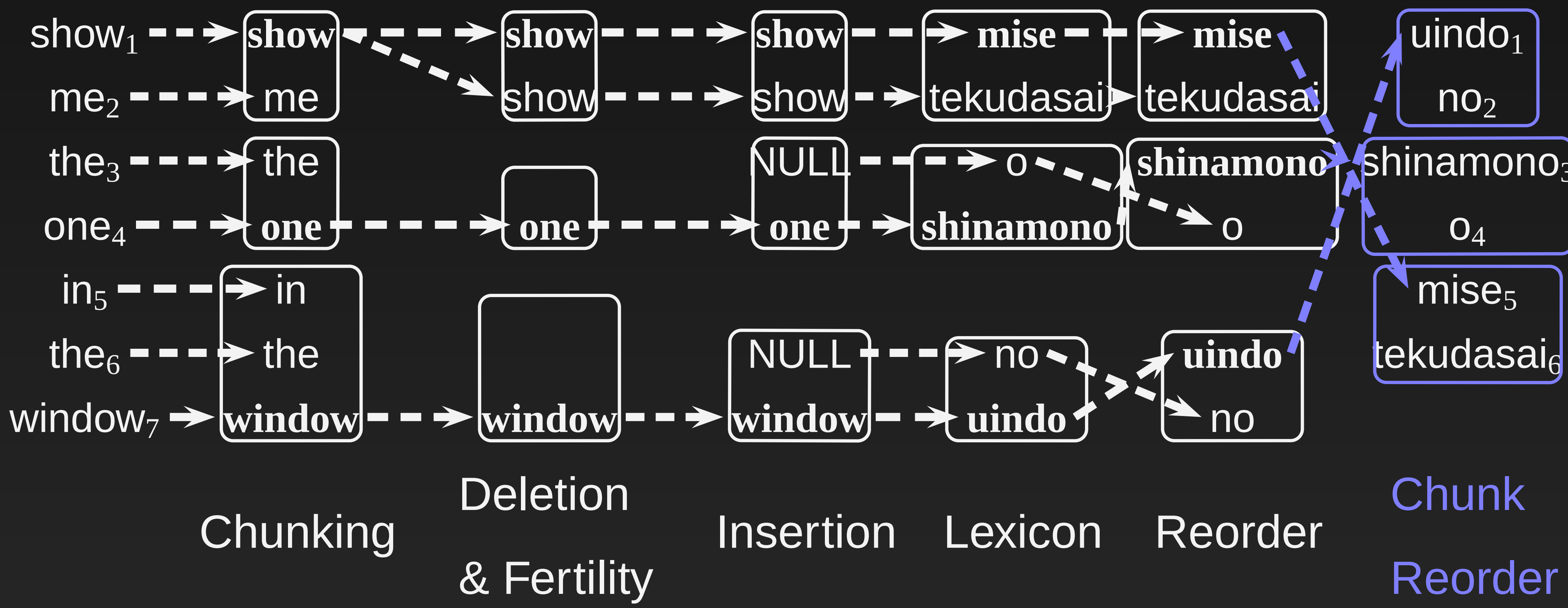
Model Structure



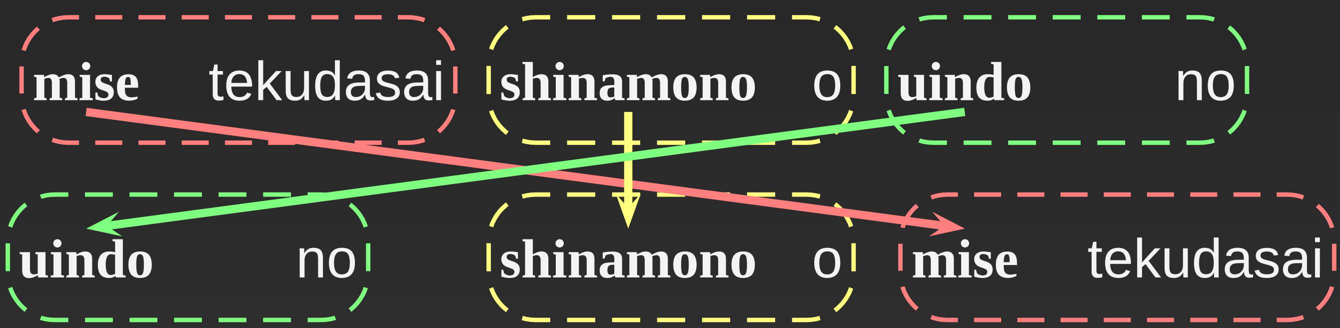
Reorder — $\prod_j P(\mathcal{A}_j | \mathcal{E}_{A_j}, \mathcal{J}_j)$



Model Structure



Chunk Reorder — $P(A|\mathcal{E}, \mathcal{J})$



Characteristics

- String-to-String translation model with a hidden chunk layer

Characteristics

- String-to-String translation model
  with a hidden chunk layer
- Bag-of-words to bag-of-words translation
  - Chunking – Translate – Reorder

Characteristics

- String-to-String translation model
  with a hidden chunk layer
- Bag-of-words to bag-of-words translation
  - Chunking – Translate – Reorder
- Chunk-wise word insertion
  vs. Sentence-wise insertion

Characteristics

- String-to-String translation model with a hidden chunk layer
- Bag-of-words to bag-of-words translation
  - Chunking – Translate – Reorder
- Chunk-wise word insertion vs. Sentence-wise insertion
- Chunking/Translate/Reorder by hypothesized “head” words

Parameter Estimation

- EM-Algorithm
- E-step: for each pair E and J

$$P(\mathcal{J}, A, \mathcal{A}, \mathcal{E}|J, E) = \frac{P(J, \mathcal{J}, A, \mathcal{A}, \mathcal{E}|E)}{\sum_{\mathcal{J}, A, \mathcal{A}, \mathcal{E}} P(J, \mathcal{J}, A, \mathcal{A}, \mathcal{E}|E)}$$

  Then, computes expectation

- M-step: From expectation, induce parameters

Some Tricks

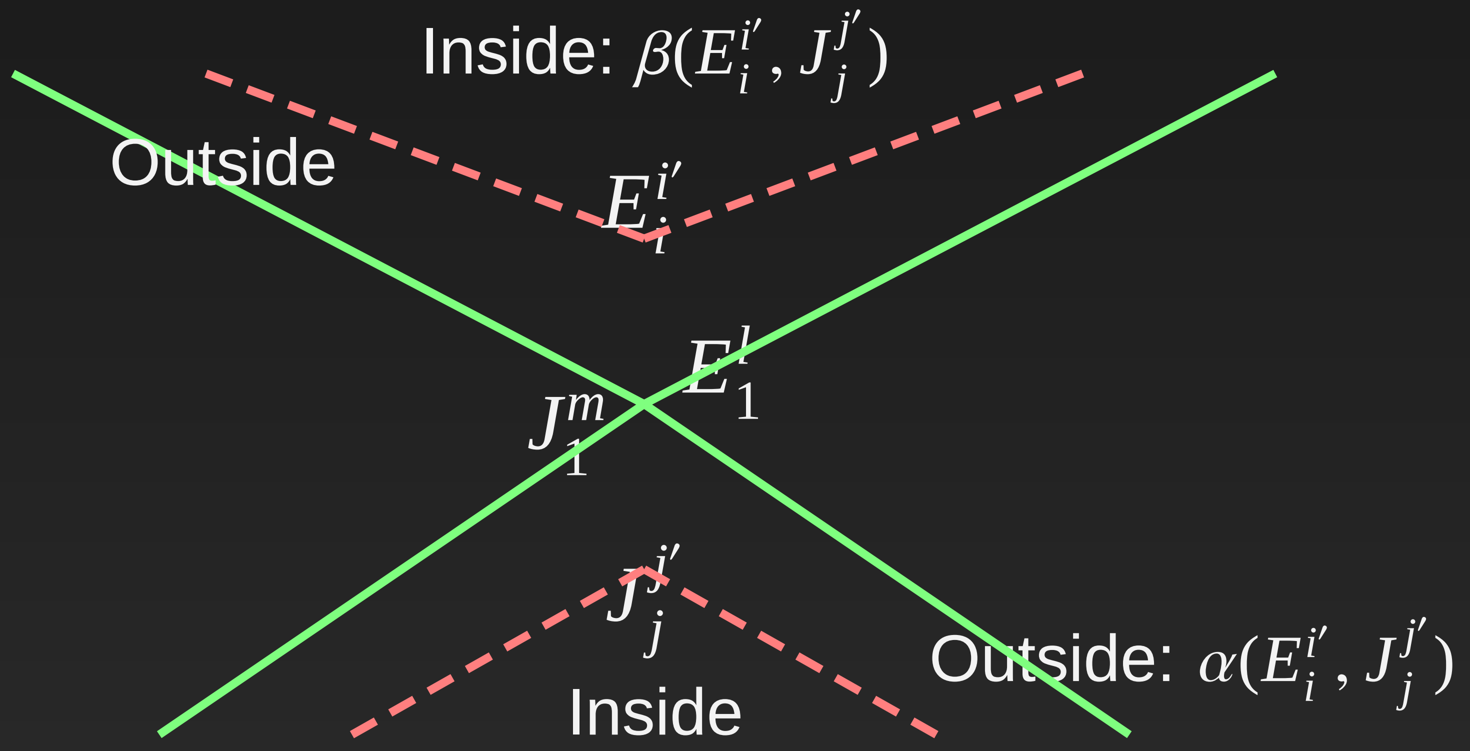
- Computational problem
- Local maximum problem

Some Tricks

- Computational problem
  - Inside-Outside Algorithm
  - Approximation
- Local maximum problem

Some Tricks

- Computational problem
  - Inside-Outside Algorithm



  - Approximation
- Local maximum problem

Some Tricks

- Computational problem
  - Inside-Outside Algorithm
  - Approximation
    - All possible word alignment: $O(lmk^4(k+1)^k))$
    - All possible chunking/alignment: $O(2^l 2^m n!)$
- Local maximum problem

Some Tricks

- Computational problem
  - Inside-Outside Algorithm
  - Approximation
    - All possible word alignment: $O(lmk^4(k+1)^k))$
    - All possible chunking/alignment: $O(2^l 2^m n!)$
    - Viterbi Chunking/Alignment + Neighbours
- Local maximum problem

Some Tricks

- Computational problem
  - Inside-Outside Algorithm
  - Approximation
- Local maximum problem

Some Tricks

- Computational problem
  - Inside-Outside Algorithm
  - Approximation
- Local maximum problem
  - Initial parameters from IBM Model 4
  - Smoothing

Decoding

- Left-to-right generation breadth-first beam search
  - Generate possible output chunks for all possible input chunks
  - Generate hypothesized output by consuming input chunks in arbitrary order and combining possible output chunks in left-to-right order

Decoding

- Left-to-right generation breadth-first beam search
  - Generate possible output chunks for all possible input chunks
  - Generate hypothesized output by consuming input chunks in arbitrary order and combining possible output chunks in left-to-right order
- Pruning
  - Beam size pruning
  - Example-based scoring

$$\log P_{tm}(J|E) + \log P_{lm}(E) + weight \times \sum_j freq(\mathcal{E}_{A_j}, \mathcal{J}_j)$$

    - Chunk-based translation model is a deficient model
    - Many model components

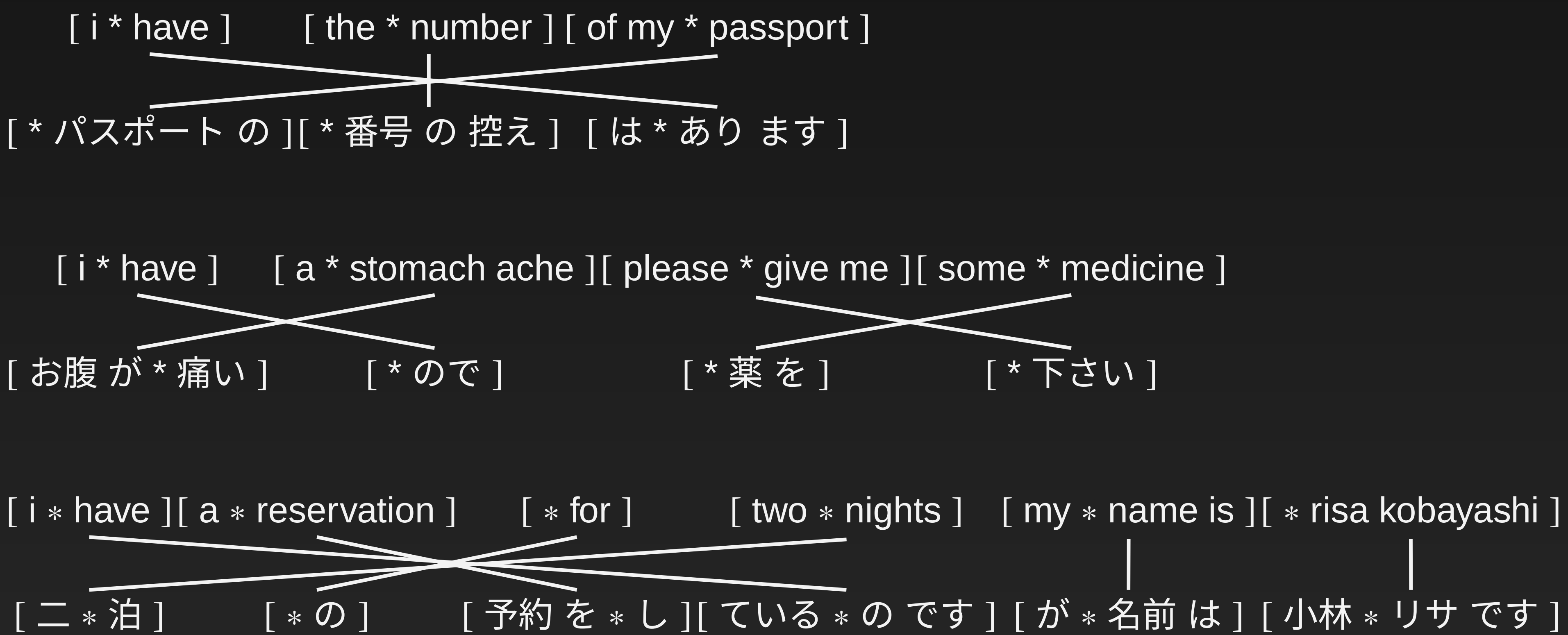
Japanese-to-English Translation Experiments

Basic Travel Expression Corpus

	Japanese	English
# of sentences	171,894	
# of words	1,181,188	1,009,065
vocabulary size	20472	16232
# of singletons	82,06	5,854
3-gram perplexity	23.7	35.8

- model4: IBM Model 4
- chunk3: Chunk-based Statistical Translation (chunk size ≤ 3)
- chunk3+: + Example-based scoring

Sample Viterbi Chunking/Alignment



Evaluation

WER: Word-error-rate, which penalizes the edit distance against reference translations.

PER: Position independent WER, which penalizes without considering positional disfluencies.

BLEU: BLEU score, which computes the ratio of n-gram for the translation results found in reference translations.

SE: Subjective evaluation ranks ranging from A to D (A:Perfect, B:Fair, C:Acceptable and D:Nonsense), judged by native speakers.

- Tested on 510 sentences
- 16 set of references for non-subjective evaluations

Results

Model	WER [%]	PER [%]	BLEU [%]	SE [%] A	SE [%] A+B	SE [%] A+B+C
model4	43.3	37.2	46.5	59.2	74.1	80.2
chunk3	40.9	36.1	48.4	59.8	73.5	78.8
chunk3+	38.5	33.7	52.1	65.1	76.3	80.6

Results

Model	WER [%]	PER [%]	BLEU [%]	SE [%] A	SE [%] A+B	SE [%] A+B+C
model4	43.3	37.2	46.5	59.2	74.1	80.2
chunk3	40.9	36.1	48.4	59.8	73.5	78.8
chunk3+	38.5	33.7	52.1	65.1	76.3	80.6

Results

Model	WER [%]	PER [%]	BLEU [%]	SE [%] A	SE [%] A+B	SE [%] A+B+C
model4	43.3	37.2	46.5	59.2	74.1	80.2
chunk3	40.9	36.1	48.4	59.8	73.5	78.8
chunk3+	38.5	33.7	52.1	65.1	76.3	80.6

Results

Model	WER [%]	PER [%]	BLEU [%]	SE [%] A	SE [%] A+B	SE [%] A+B+C
model4	43.3	37.2	46.5	59.2	74.1	80.2
chunk3	40.9	36.1	48.4	59.8	73.5	78.8
chunk3+	38.5	33.7	52.1	65.1	76.3	80.6

Sample Translations

input:	一 五 二 便 の 荷物 は これ で 全部 です か
reference:	is this all the baggage from flight one five two
model4:	is this all you baggage for flight one five two
chunk3+:	is this all the baggage from flight one five two
input:	朝食 を ルームサービス で お 願い し ます
reference:	may i have room service for breakfast please
model4:	please give me some room service please
chunk3+:	i ’d like room service for breakfast
input:	もしもし 三 月 十 九 日 の 予約 を 変更 し たい の です が
reference:	hello i ’d like to change my reservation for march nineteenth
model4:	i ’d like to change my reservation for ninety days be march hello
chunk3+:	hello i ’d like to change my reservation on march nineteenth
input:	二 三 分 待っ て下さい 今 電話 中 な ん です
reference:	wait a couple of minutes i ’m telephoning now
model4:	is this the line is busy now a few minutes
chunk3+:	i ’m on another phone now please wait a couple of minutes

Summary

- String-to-String translation model with hidden chunks
- More hidden variables
  $\longrightarrow$ More cost for training + decoding
  - Trainin Cost $\approx$ IBM Model 5 with pegging
  - Decoding Cost: moderate with Example-based scoring
- Quality Improvement: Slightly, but (probably) significant

Summary

- String-to-String translation model with hidden chunks
- More hidden variables
  $\longrightarrow$ More cost for training + decoding
  - Trainin Cost $\approx$ IBM Model 5 with pegging
  - Decoding Cost: moderate with Example-based scoring
- Quality Improvement: Slightly, but (probably) significant
- Other approaches?

Typology of Statistical Machine Translation

- Approach 1: Precomputation of Structure
- Approach 2: Structure-to-String
- Approach 3: Collection of Hierarchical FST

Typology of Statistical Machine Translation

- Approach 1: Precomputation of Structure
- Approach 2: Structure-to-String
- Approach 3: Collection of Hierarchical FST

Typology of Statistical Machine Translation

- Approach 1: Precomputation of Structure
  - Templates (Och et al. 1999)
  - Chunks from syntax-based phrase alignment (Watanabe et al. 2002)
  - Direct phrase induction (Marcu and Wong 2002)
    - Bias the training corpus by template, chunk or phrase
    - Works significantly better on observed word sequences, but not for unseen sequences
- Approach 2: Structure-to-String
- Approach 3: Collection of Hierarchical FST

Typology of Statistical Machine Translation

- Approach 1: Precomputation of Structure
- Approach 2: Structure-to-String
- Approach 3: Collection of Hierarchical FST

Typology of Statistical Machine Translation

- Approach 1: Precomputation of Structure
- Approach 2: Structure-to-String
  - Phrase-to-string Modeling (Wang 1998)
  - Syntax-to-string Modeling (Yamada and Knight 2001)
    - Bias the source part of a training corpus by “structure”
    - Computationally cheaper
    - Relies on the monolingual processing (parser or chunker)
- Approach 3: Collection of Hierarchical FST

Typology of Statistical Machine Translation

- Approach 1: Precomputation of Structure
- Approach 2: Structure-to-String
- Approach 3: Collection of Hierarchical FST

Typology of Statistical Machine Translation

- Approach 1: Precomputation of Structure
- Approach 2: Structure-to-String
- Approach 3: Collection of Hierarchical FST
  - (Alshawi et al. 2000)
    - Deterministic vs. Non-Deterministic
    - Faster decoding + less space
      vs. Slow decoding + pruning
    - Limited domain vs. Larger domain