Left-to-Right Target Generation for Hierarchical Phrase-based Translation

Marzieh Razavi

Maryam Siahbani

Ravikiran Vadlapudi

Introduction

- Problem with Hierarchical phrase-based translation: **Cost of Decoding**
  - the number of extracted rules from initial phrases would be numerous
  - The integration with language model especially when incorporating with higher order n-grams:
  - $O\left(n^3|T|^{4(m-1)}\right)$ (T: English terminal alphabet)

Introduction

- Idea : **Target-normalized** Hierarchical Phrase-based Translation
  - Restrict the target side to have GNF-like structure:
    - Sequence of terminals followed by non-terminals

➡ Reduce the number of extracted rules from the bilingual corpus

➡ Integration with n-gram language model would be straight forward

- The target side is generated in Left-to-right order

Translation Model

$$\boldsymbol{X \rightarrow < \gamma,\ \bar{b}\beta, \sim >}$$

- X : non-terminal
- γ : strings of terminals and non-terminals for source
- $\bar{b}$: strings of terminals or phrase for target
- β : string of non-terminals for target
- $\sim$: 1-1 correspondence between non-terminals

Rule Extraction

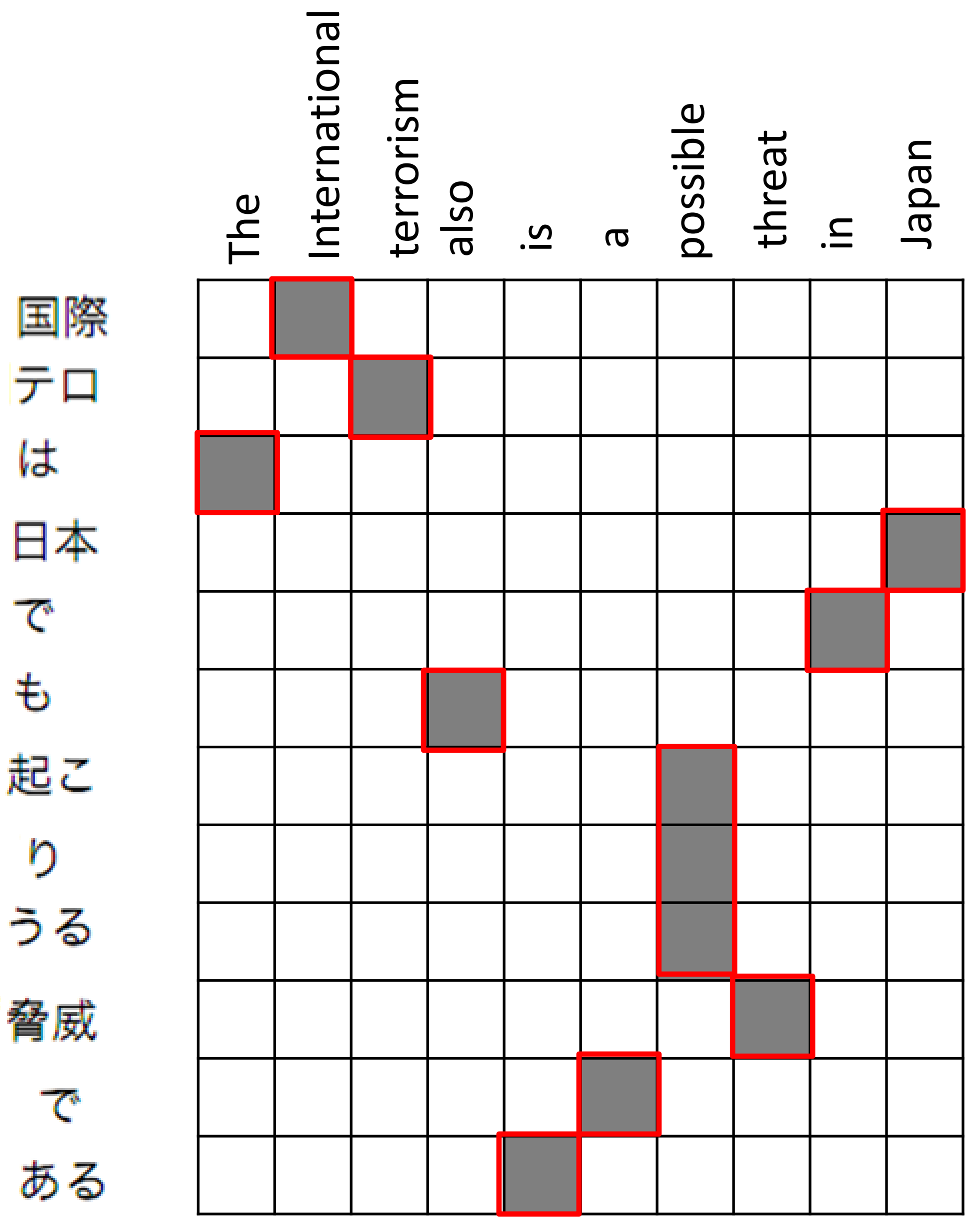
1. Identifying initial phrase pairs

2. Extracting rules:
   I. Use initial phrases $(\bar{\boldsymbol{f}}, \bar{\boldsymbol{e}})$:

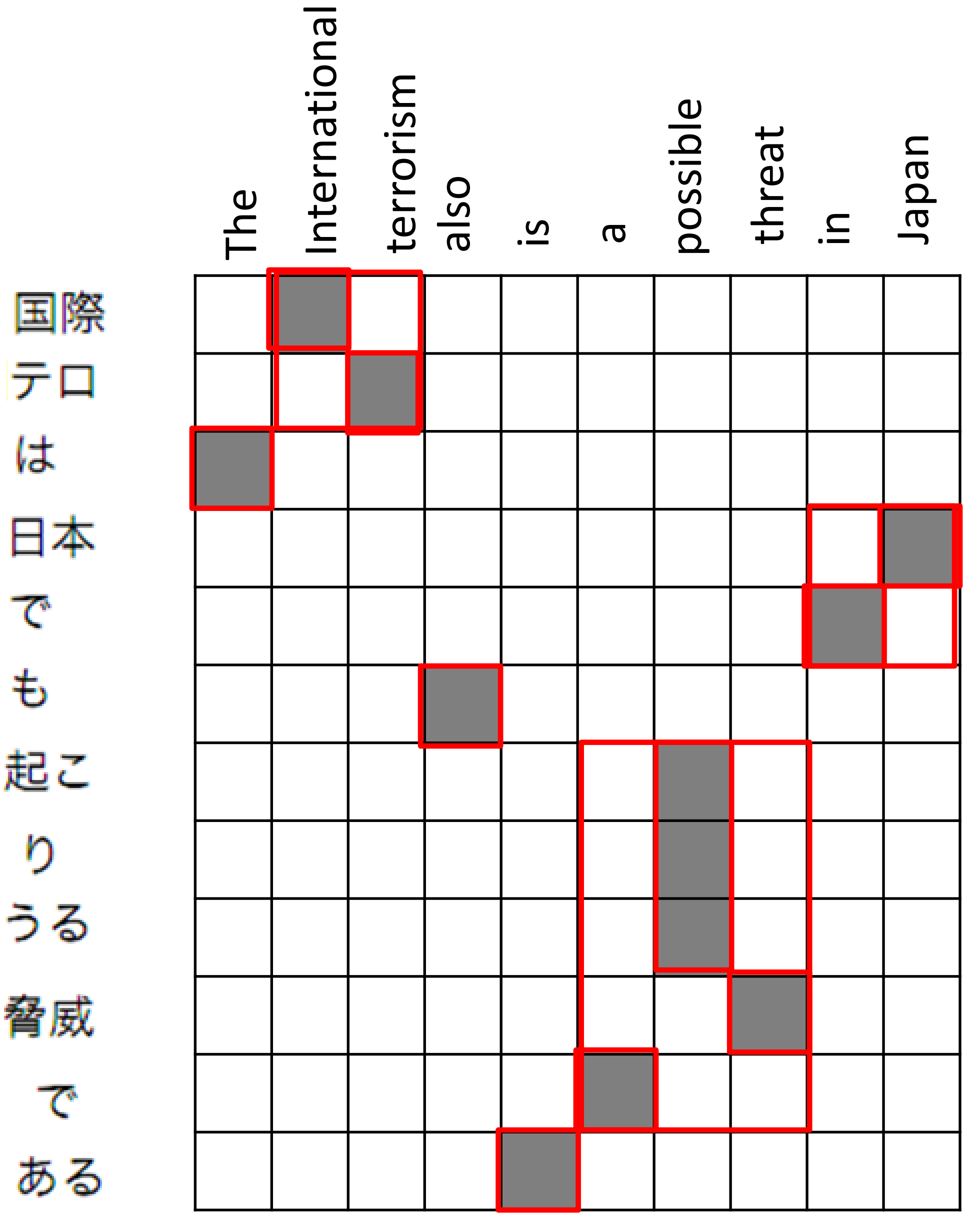
$$\boldsymbol{X} \rightarrow < \bar{\boldsymbol{f}}, \bar{\boldsymbol{e}} >$$

   II. Choose a rule $\boldsymbol{X} \rightarrow < \boldsymbol{\gamma}, \boldsymbol{\alpha} >$ and a phrase $(\bar{\boldsymbol{f}}, \bar{\boldsymbol{e}})$ s.t. $\gamma = \gamma' \bar{f} \gamma''$ and $\alpha = \bar{e'} \bar{e} \beta$

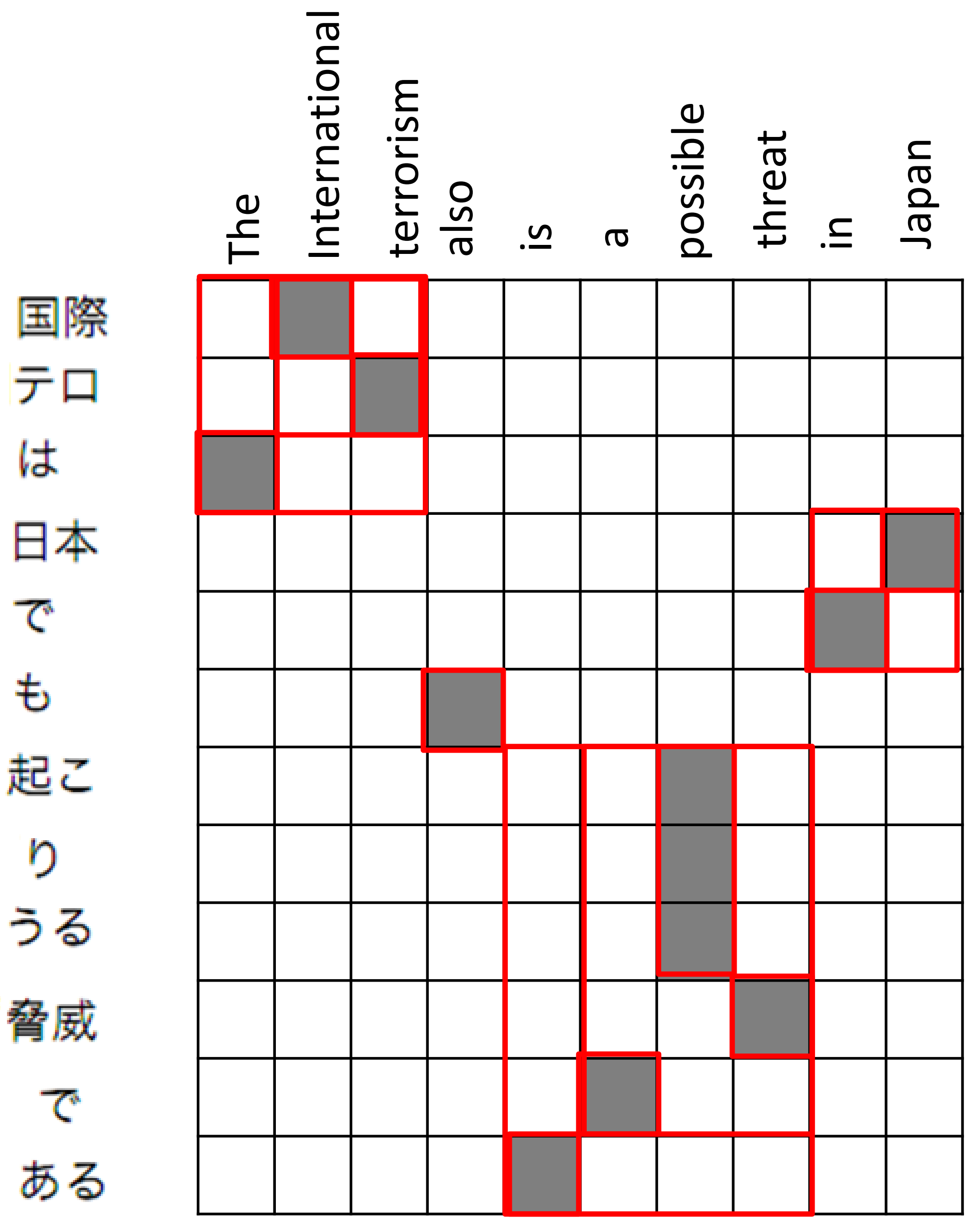
$$\boldsymbol{X} \rightarrow < \gamma' \underset{k}{X} \gamma'', \bar{e'} \underset{k}{X} \beta >$$

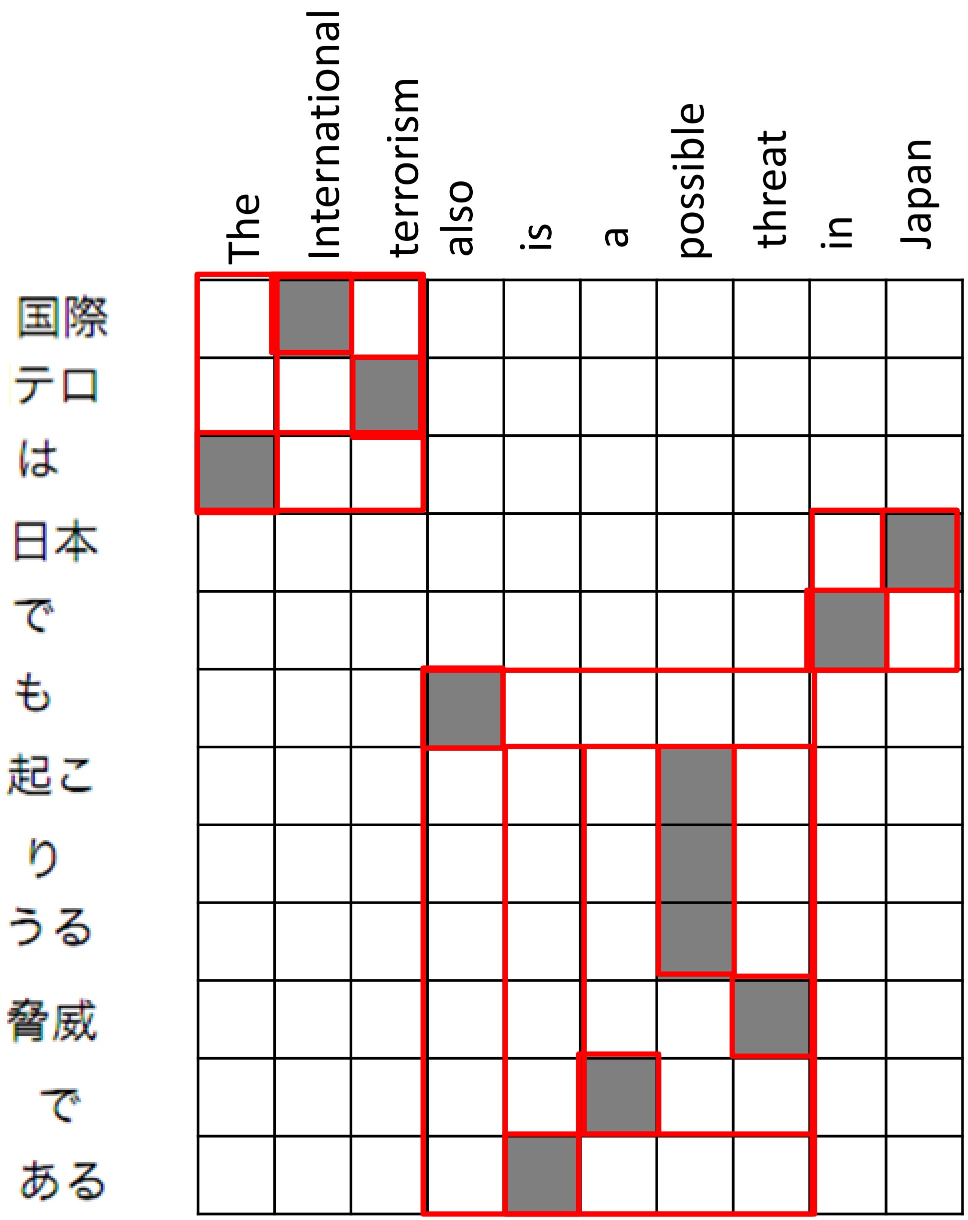


The
International
terrorism
also
is
a
possible
threat
in
Japan
国際
テロ
は
日本
で
も
起こ
り
うる
脅威
で
ある

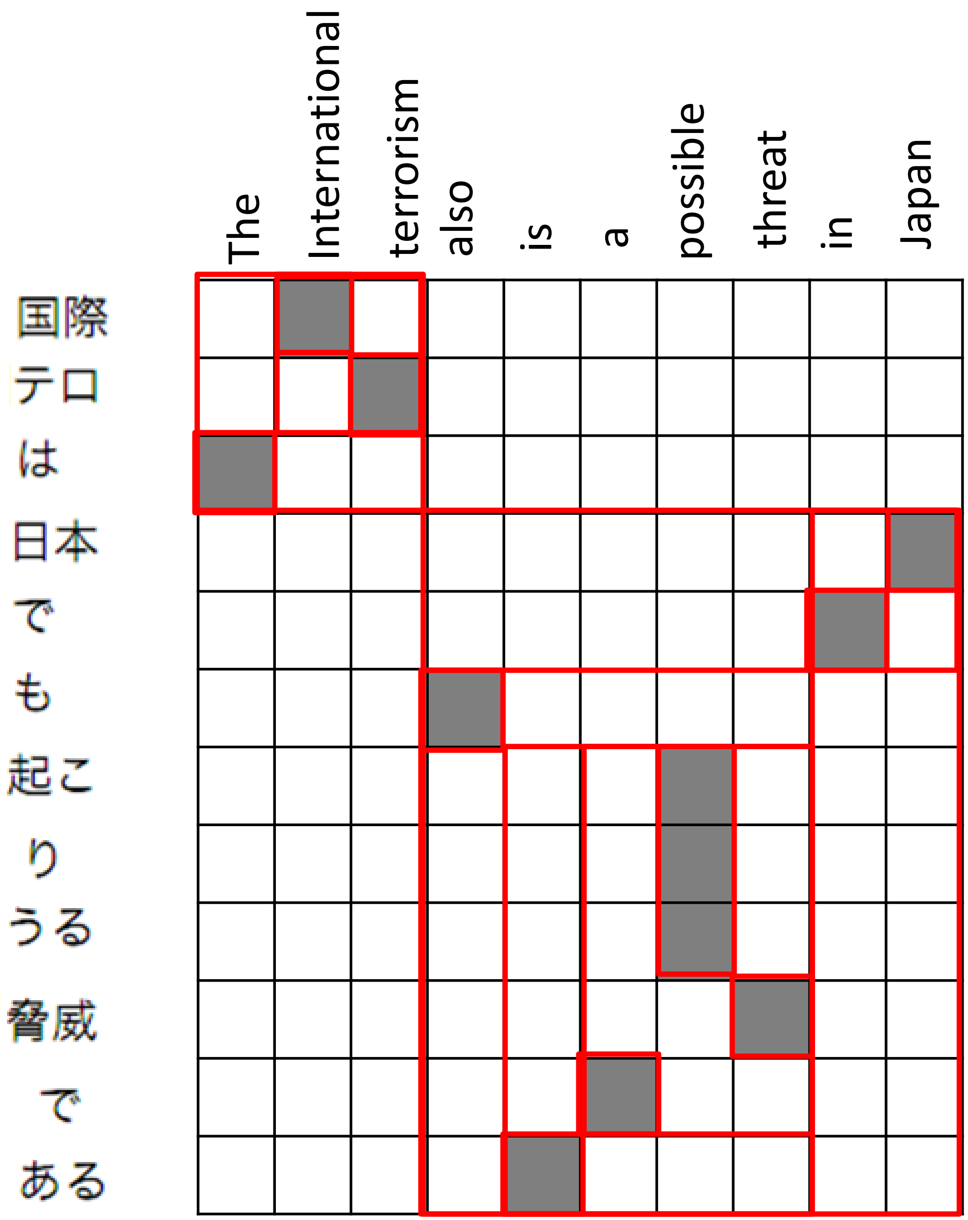


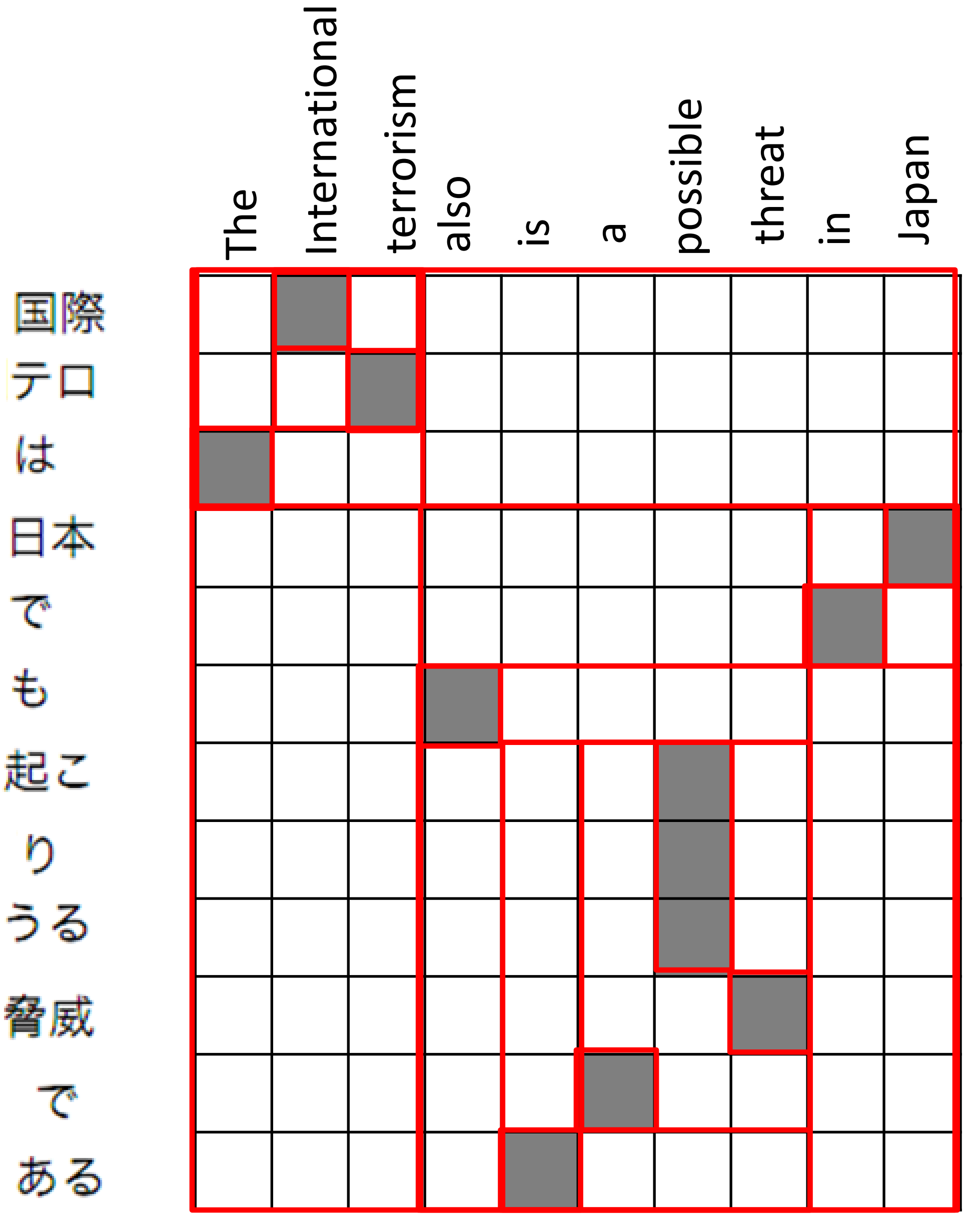
The
International
terrorism
also
is
a
possible
threat
in
Japan
国際
テロ
は
日本
で
も
起こ
り
うる
脅威
で
ある

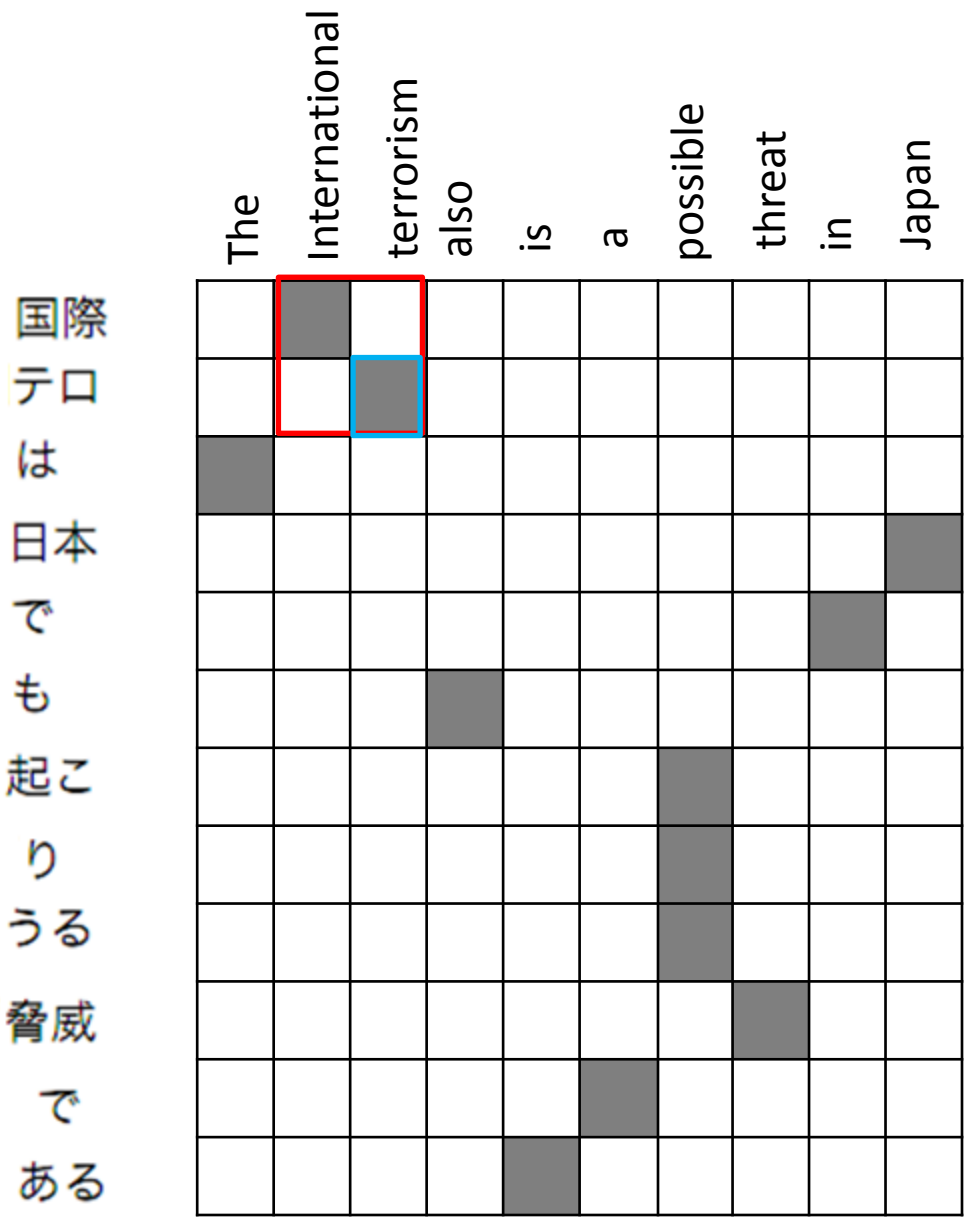

The
International
terrorism
also
is
a
possible
threat
in
Japan
国際
テロ
は
日本
で
も
起こ
り
うる
脅威
で
ある



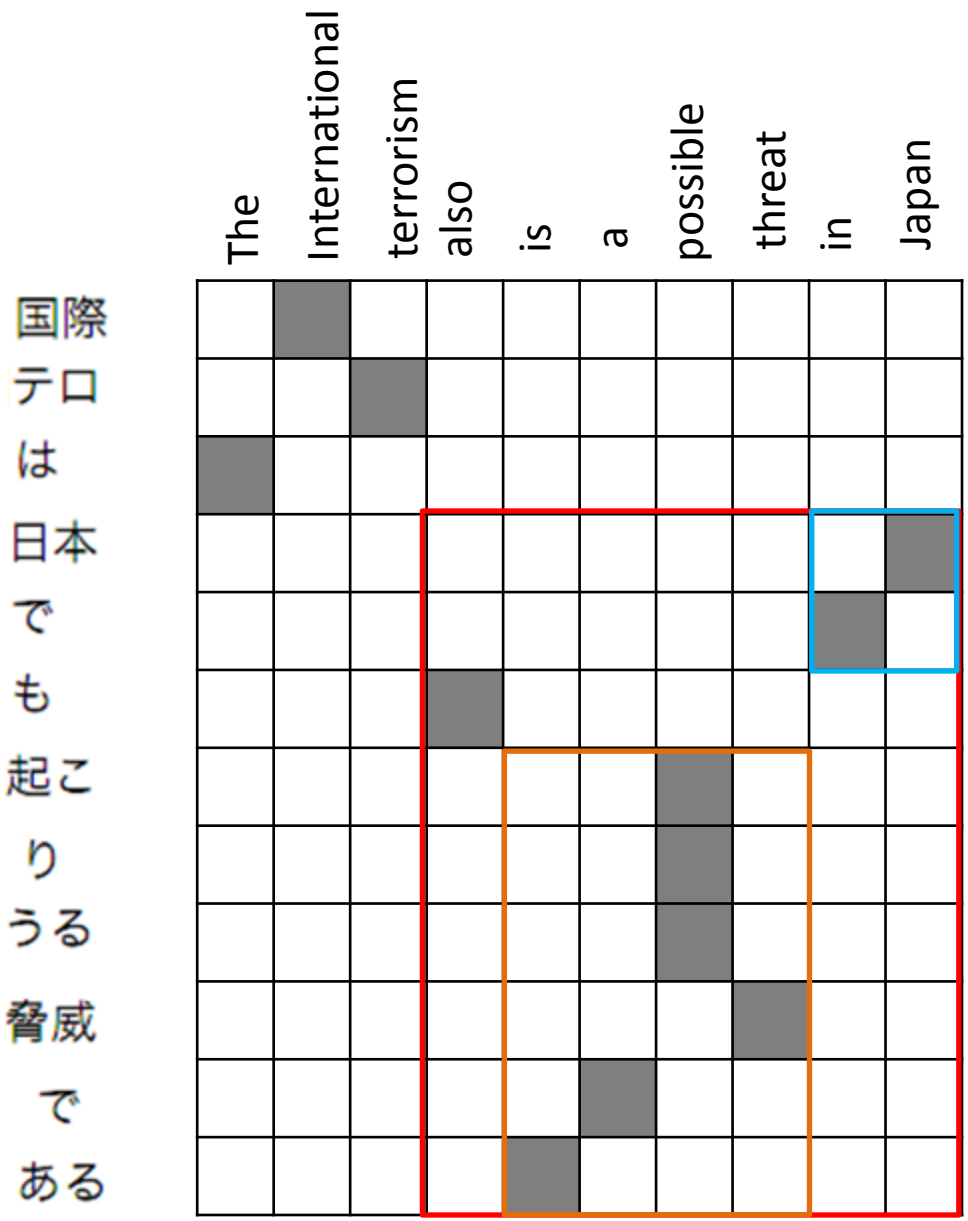
The
International
terrorism
also
is
a
possible
threat
in
Japan
国際
テロ
は
日本
で
も
起こ
り
うる
脅威
で
ある


The
International
terrorism
also
is
a
possible
threat
in
Japan
国際
テロ
は
日本
で
も
起こ
り
うる
脅威
で
ある


The
International
terrorism
also
is
a
possible
threat
in
Japan
国際
テロ
は
日本
で
も
起こ
り
うる
脅威
で
ある



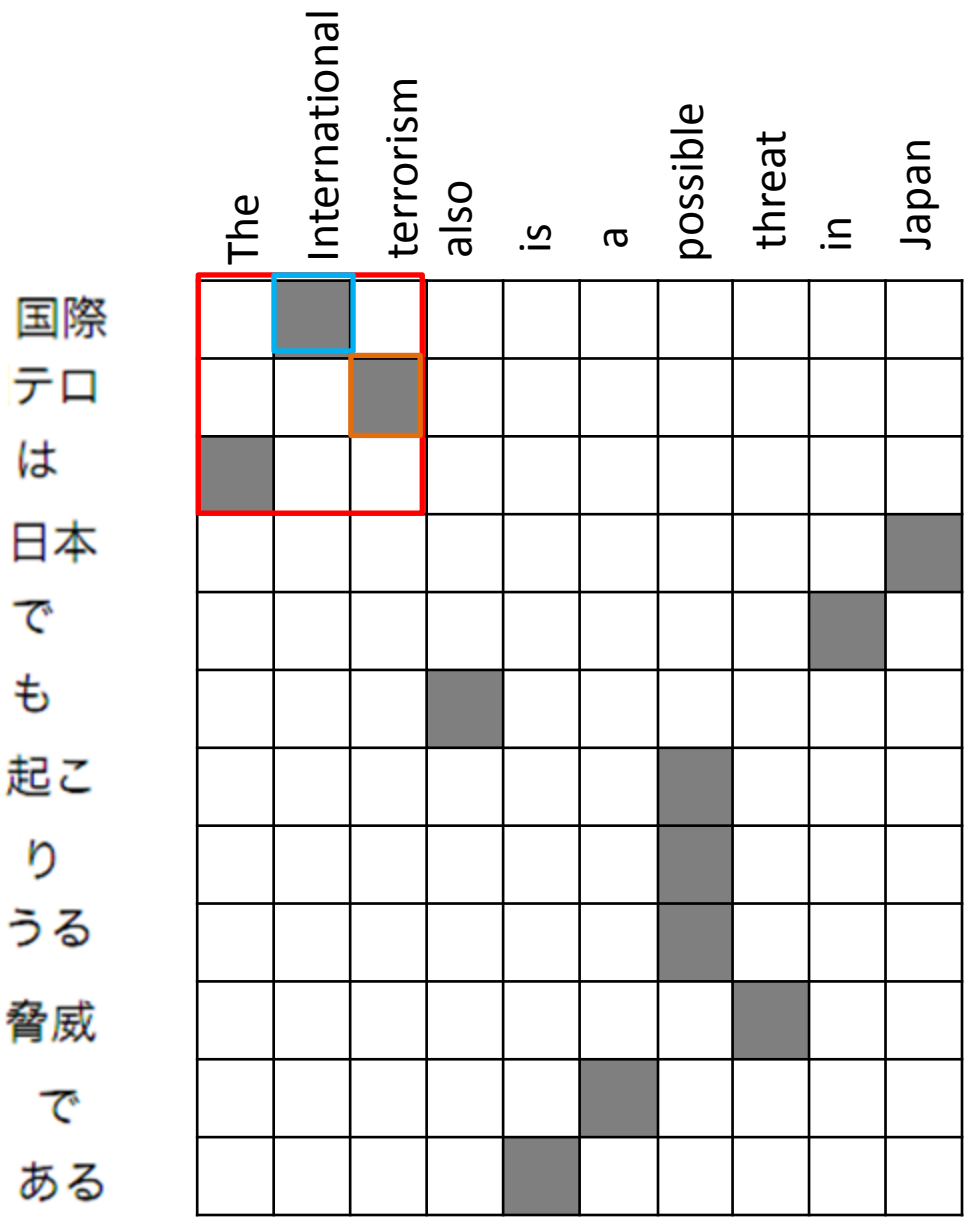
$$X \rightarrow < \text{国際} \; X_{[1]}, int.X_{[1]} >$$



$X \rightarrow <X$ も 起こり うる 脅威 で ある $>$ [1]

$, also\ is\ a\ possible\ treat\ X >$ [1]

$X \rightarrow <X$ も X [1] [2], $also\ X$ [2] $X>$ [1]

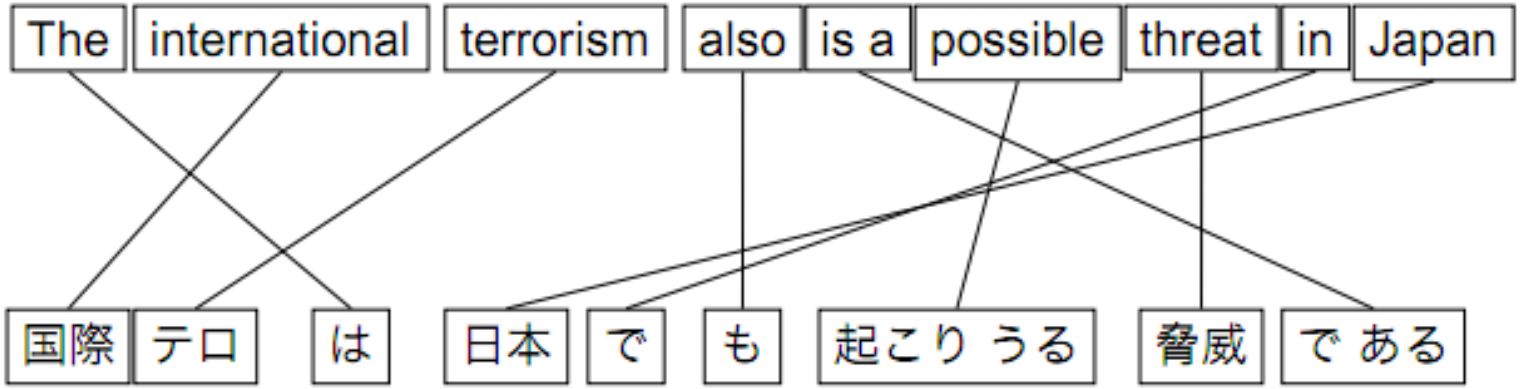


$$X \rightarrow < X_{[1]} \text{は} X_{[2]}, the\ X_{[1]}\ X_{[2]} >$$

Phrase-based Rules

- From step 1 of rule extraction : $X \rightarrow < \bar{f}, \bar{e} >$
- We add the following rules to avoid data sparseness:
  - $X \rightarrow < \bar{f}\ X_1, \bar{e}\ X_1 >$
  - $X \rightarrow < X_1 \bar{f}, \bar{e}\ X_1 >$
  - $X \rightarrow < X_1 \bar{f}\ X_2, \bar{e}\ X_1 X_2 >$
  - $X \rightarrow < X_2 \bar{f}\ X_1, \bar{e} X_1 X_2 >$

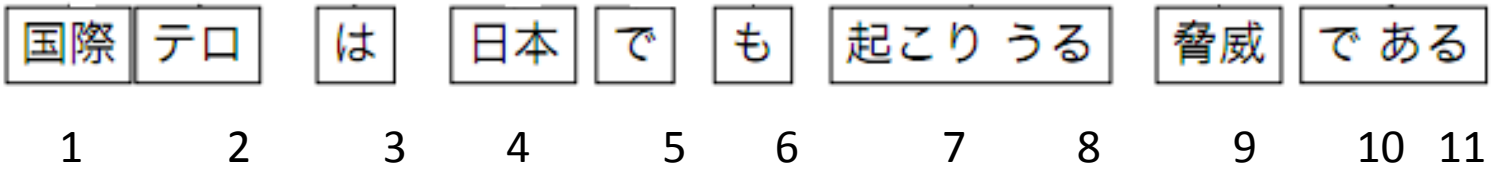
Example


The
international
terrorism
also
is a
possible
threat
in
Japan
国際
テロ
は
日本
で
も
起こり うる
脅威
で ある

Rules | Stack

[1, 11]

$X : [1, 11] \rightarrow \langle X_{[1]} : [1, 2]$ は $X_{[2]} : [4, 11]$, The $X_{[1]}\ X_{[2]} \rangle$

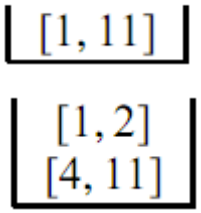


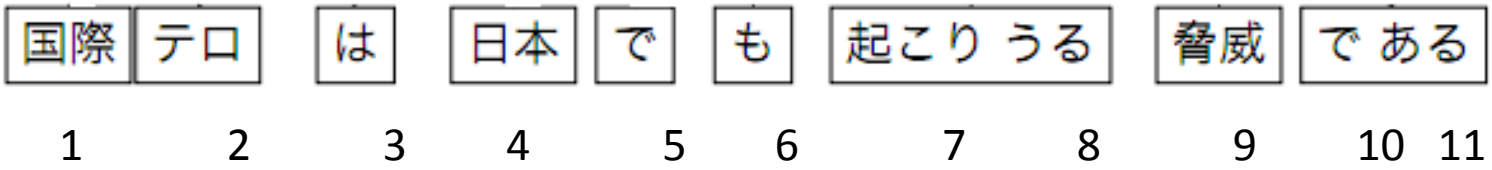
Rules

Stack

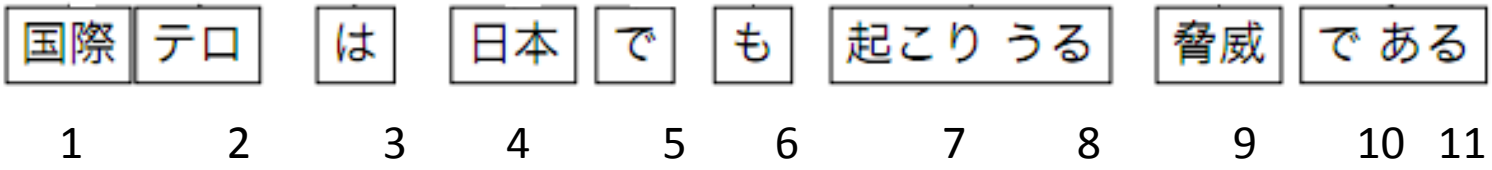
$X : [1, 11] \rightarrow \langle X_{\boxed{1}} : [1, 2]$ は $X_{\boxed{2}} : [4, 11]$, The $X_{\boxed{1}}$ $X_{\boxed{2}} \rangle$

$X : [1, 2] \rightarrow \langle$ 国際 $X_{\boxed{1}} : [2, 2]$, international $X_{\boxed{1}} \rangle$

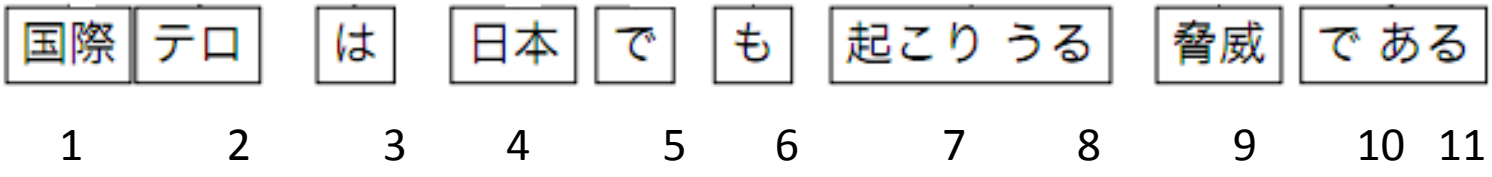




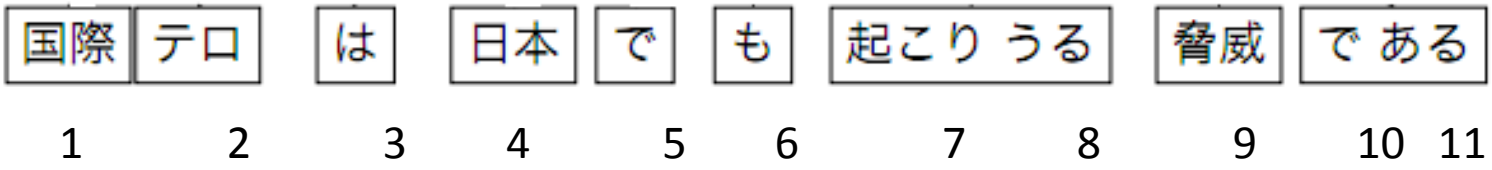
Rules	Stack
	[1, 11]
$X : [1, 11] \rightarrow \langle X_{\boxed{1}} : [1, 2]$ は $X_{\boxed{2}} : [4, 11]$, The $X_{\boxed{1}}$ $X_{\boxed{2}} \rangle$	[1, 2] [4, 11]
$X : [1, 2] \rightarrow \langle$国際 $X_{\boxed{1}} : [2, 2]$, international $X_{\boxed{1}} \rangle$	[2, 2] [4, 11]
$X : [2, 2] \rightarrow \langle$テロ, terrorism$\rangle$	[4, 11]



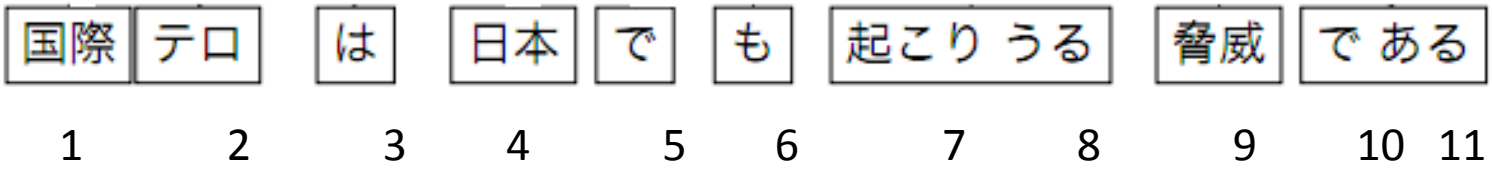
Rules	Stack
	[1, 11]
$X : [1, 11] \rightarrow \langle X_{\boxed{1}} : [1, 2]$ は $X_{\boxed{2}} : [4, 11]$, The $X_{\boxed{1}}$ $X_{\boxed{2}} \rangle$	[1, 2] [4, 11]
$X : [1, 2] \rightarrow \langle$国際 $X_{\boxed{1}} : [2, 2]$, international $X_{\boxed{1}} \rangle$	[2, 2] [4, 11]
$X : [2, 2] \rightarrow \langle$テロ, terrorism$\rangle$	[4, 11]
$X : [4, 11] \rightarrow \langle X_{\boxed{2}} : [4, 5]$ も $X_{\boxed{1}} : [7, 11]$, also $X_{\boxed{1}}$ $X_{\boxed{2}} \rangle$	[7, 11] [4, 5]



Rules	Stack
	[1, 11]
$X : [1, 11] \rightarrow \langle X_{[1]} : [1, 2]$ は $X_{[2]} : [4, 11]$, The $X_{[1]}$ $X_{[2]} \rangle$	[1, 2] [4, 11]
$X : [1, 2] \rightarrow \langle$国際 $X_{[1]} : [2, 2]$, international $X_{[1]} \rangle$	[2, 2] [4, 11]
$X : [2, 2] \rightarrow \langle$テロ, terrorism$\rangle$	[4, 11]
$X : [4, 11] \rightarrow \langle X_{[2]} : [4, 5]$ も $X_{[1]} : [7, 11]$, also $X_{[1]}$ $X_{[2]} \rangle$	[7, 11] [4, 5]
$X : [7, 11] \rightarrow \langle X_{[1]} : [7, 9]$ で ある, is a $X_{[1]} \rangle$	[7, 9] [4, 5]



Rules	Stack
$X : [7, 11] \rightarrow \langle X_{\boxed{1}} : [7, 9]$ で ある, is a $X_{\boxed{1}} \rangle$	[7, 9] [4, 5]
$X : [7, 9] \rightarrow \langle$起こり うる $X_{\boxed{1}} : [9, 9]$, possible $X_{\boxed{1}} \rangle$	[9, 9] [4, 5]



Rules	Stack
$X : [7, 11] \rightarrow \langle X_{\boxed{1}} : [7, 9]$ で ある, is a $X_{\boxed{1}} \rangle$	[7, 9] [4, 5]
$X : [7, 9] \rightarrow \langle$起こり うる $X_{\boxed{1}} : [9, 9]$, possible $X_{\boxed{1}} \rangle$	[9, 9] [4, 5]
$X : [9, 9] \rightarrow \langle$脅威, threat$\rangle$	[4, 5]

Rules	Stack
$X : [7, 11] \rightarrow \langle X_{[1]} : [7, 9]$ で ある, is a $X_{[1]} \rangle$	[7, 9] [4, 5]
$X : [7, 9] \rightarrow \langle$起こり うる $X_{[1]} : [9, 9]$, possible $X_{[1]} \rangle$	[9, 9] [4, 5]
$X : [9, 9] \rightarrow \langle$脅威, threat$\rangle$	[4, 5]
$X : [4, 5] \rightarrow \langle X_{[1]} : [4, 4]$ で, in $X_{[1]} \rangle$	[4, 4]

Rules	Stack
$X : [7, 11] \rightarrow \langle X_{\boxed{1}} : [7, 9]$ で ある, is a $X_{\boxed{1}} \rangle$	[7, 9] [4, 5]
$X : [7, 9] \rightarrow \langle$起こり うる $X_{\boxed{1}} : [9, 9]$, possible $X_{\boxed{1}} \rangle$	[9, 9] [4, 5]
$X : [9, 9] \rightarrow \langle$脅威, threat$\rangle$	[4, 5]
$X : [4, 5] \rightarrow \langle X_{\boxed{1}} : [4, 4]$ で, in $X_{\boxed{1}} \rangle$	[4, 4]
$X : [4, 4] \rightarrow \langle$日本, Japan$\rangle$	

$X : [1, 11] \rightarrow \langle X_{[1]} : [1, 2]$ は $X_{[2]} : [4, 11]$, The $X_{[1]}\ X_{[2]} \rangle$

$X : [1, 2] \rightarrow \langle$国際 $X_{[1]} : [2, 2]$, international $X_{[1]} \rangle$

$X : [2, 2] \rightarrow \langle$テロ, terrorism$\rangle$

$X : [4, 11] \rightarrow \langle X_{[2]} : [4, 5]$ も $X_{[1]} : [7, 11]$, also $X_{[1]}\ X_{[2]} \rangle$

$X : [7, 11] \rightarrow \langle X_{[1]} : [7, 9]$ で ある, is a $X_{[1]} \rangle$

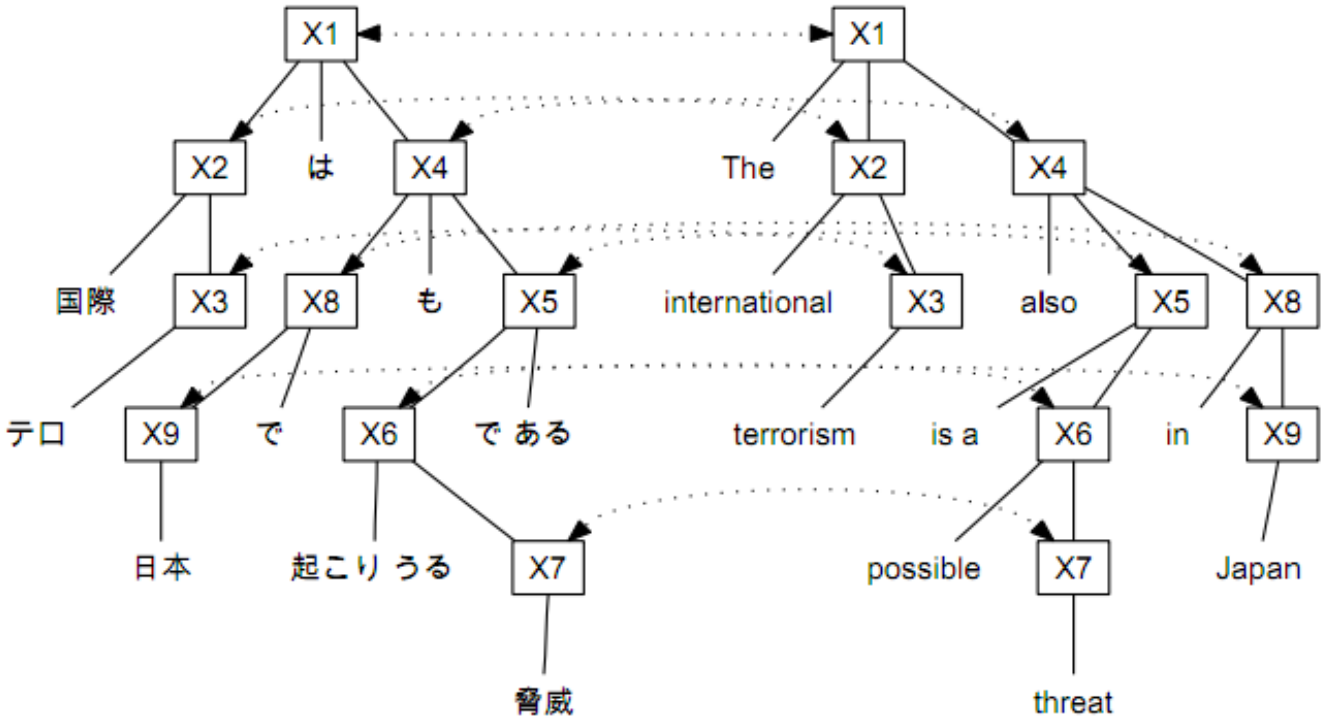
$X : [7, 9] \rightarrow \langle$起こり うる $X_{[1]} : [9, 9]$, possible $X_{[1]} \rangle$

$X : [9, 9] \rightarrow \langle$脅威, threat$\rangle$

$X : [4, 5] \rightarrow \langle X_{[1]} : [4, 4]$ で, in $X_{[1]} \rangle$

$X : [4, 4] \rightarrow \langle$日本, Japan$\rangle$

Example of derivation tree



Model

- General log-linear model over derivations D

$$\underset{e_1{}^I}{\text{argmax}} \frac{exp\left(\lambda_{lm} P_{lm}\left(e_1{}^I\right) + \sum_{m=1,m\neq lm}^{M} \lambda_m h_m\left(e_1{}^I | f_1{}^J, D\right)\right)}{\sum_{e'_1{}^{I'}} exp\left(\lambda_{lm} P_{lm}\left(e'_1{}^{I'}\right) + \sum_{m=1,m\neq lm}^{M} \lambda_m h_m\left(e'_1{}^{I'} | f_1{}^J, D\right)\right)}$$

  - $h_m\left(e_1{}^I | f_1{}^J, D\right)$: feature functions
  - λ_m: weights

Feature functions

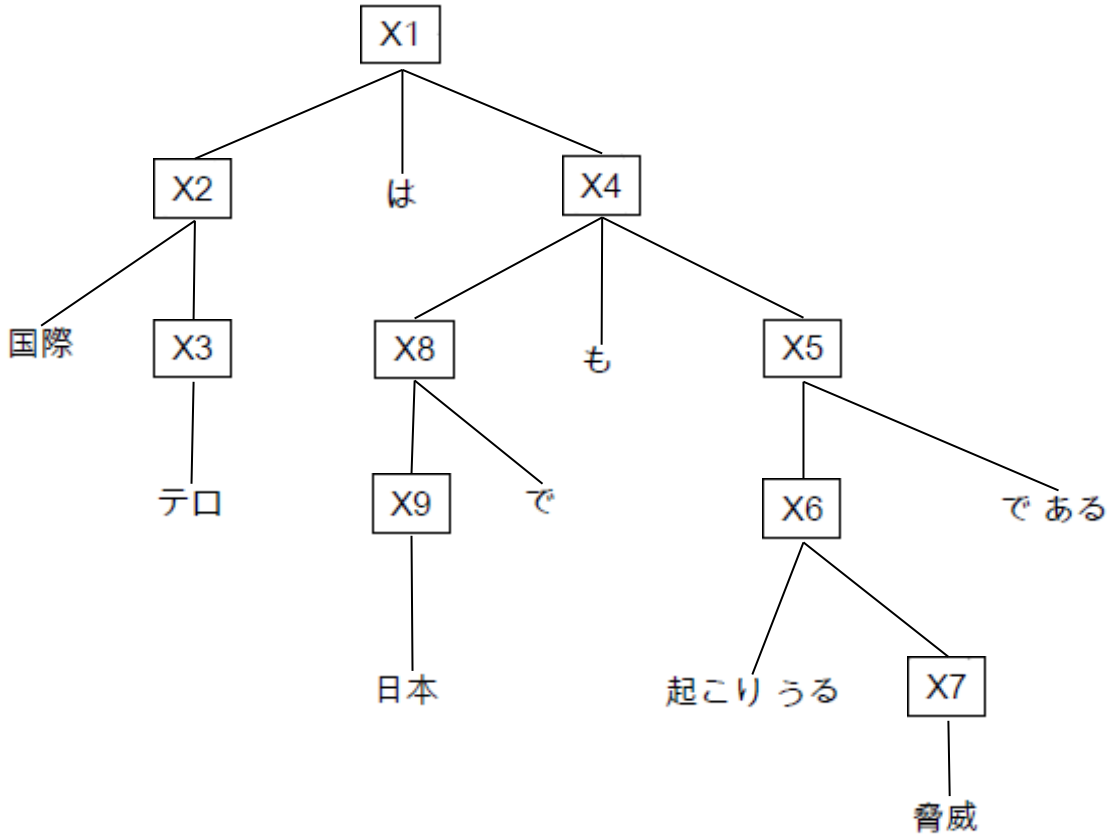
- Likelihood of two sentences ${f_1}^J$ and ${e_1}^I$
  - $h_\phi(e_1^I|f_1^J, \mathcal{D})$ ， $h_\phi(f_1^J|e_1^I, \mathcal{D})$

$$h_\phi(f_1^J|e_1^I, \mathcal{D}) = \log \prod_{\langle\gamma,\alpha\rangle\in\mathcal{D}} \phi(\gamma|\alpha)$$

- Lexical weights: $h_w({e_1}^I|{f_1}^J, D), h_w({f_1}^J|{e_1}^I, D)$
  - How well the words in ${e_1}^I$ translate the words in${f_1}^J$
- Language model

Reordering features

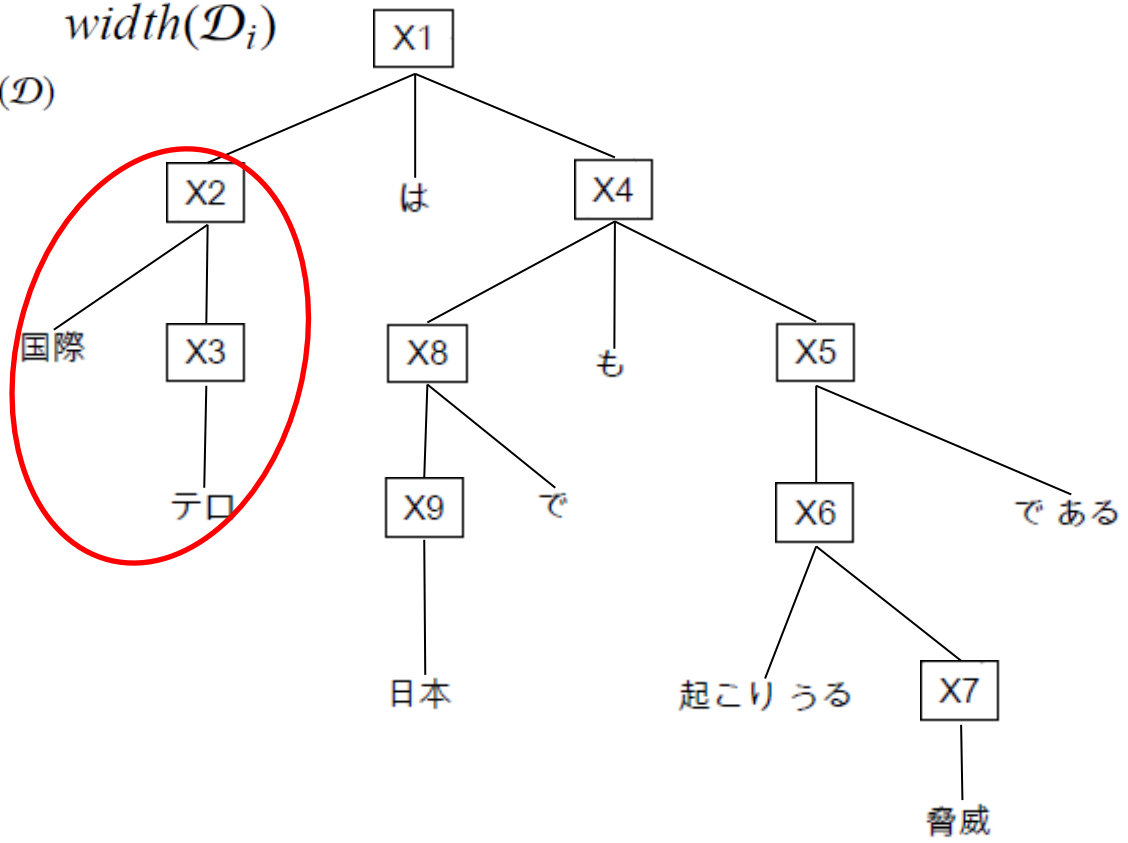
- Two features to limit the reordering



Reordering features

$$h_h(e_1^I, f_1^J, \mathcal{D}) = \sum_{\mathcal{D}_i \in back(\mathcal{D})} height(\mathcal{D}_i)$$

$$h_w(e_1^I, f_1^J, \mathcal{D}) = \sum_{\mathcal{D}_i \in back(\mathcal{D})} width(\mathcal{D}_i)$$



Length-based features

- Three length-based features:
  - Length of English sentence: $h_l(e_1{}^I) = I$
    - Control the length of output
  - No. of extracted rules: $h_r(D) = rule(D)$
  - No. of phrase-based rules: $h_p(D) = phrase(D)$
    - Control whether to choose a extracted rule of phrase-based rules in D

Experiments

	# rules/phrases
Phrase	5,433,091
Normalized-2	6,225,630
Normalized-3	6,233,294
Hierarchical	12,824,387

- No. of Rules with 2 non-terminals are slightly larger than phrase pairs
- Including 3 non-terminals did not change the grammar size

Experiments

- At most two non-terminals
- 3-gram/5-gram language models

		BLEU [%]	NIST
Phrase	3-gram	7.14	3.21
	5-gram	7.33	3.19
Normalized-2	3-gram	10.00	4.11
	5-gram	10.26	4.20

results for Japanese-English

Questions ??