

# **Example-based Machine Translation based on Deeper NLP**

Toshiaki Nakazawa<sup>1</sup>, Kun Yu<sup>1</sup>, Sadao Kurohashi<sup>2</sup>

1. Graduate School of Information Science and Technology,  
The University of Tokyo, Tokyo, Japan, 113-8656
2. Graduate School of Informatics,  
Kyoto University, Kyoto, Japan, 606-8501

# **Outline**

- **Why EBMT?**
- **Description of Kyoto-U EBMT System**
- **Japanese Particular Processing**
  - **Pronoun Estimation**
  - **Japanese Flexible Matching**
- **Result and Discussion**
- **Conclusion and Future Work**

# **Outline**

- **Why EBMT?**
- **Description of Kyoto-U EBMT System**
- **Japanese Particular Processing**
  - **Pronoun Estimation**
  - **Japanese Flexible Matching**
- **Result and Discussion**
- **Conclusion and Future Work**

# **Why EBMT?**

- **Pursuing deep NLP**
  - **Improvement of fundamental analyses leads to improvement of MT**
  - **Feedback from MT can be expected**
  
- **EBMT setting is suitable in many cases**
  - **Not a large corpus, but similar translation examples in relatively close domain**
  - **e.g. manual translation, patent translation, ...**

# Outline

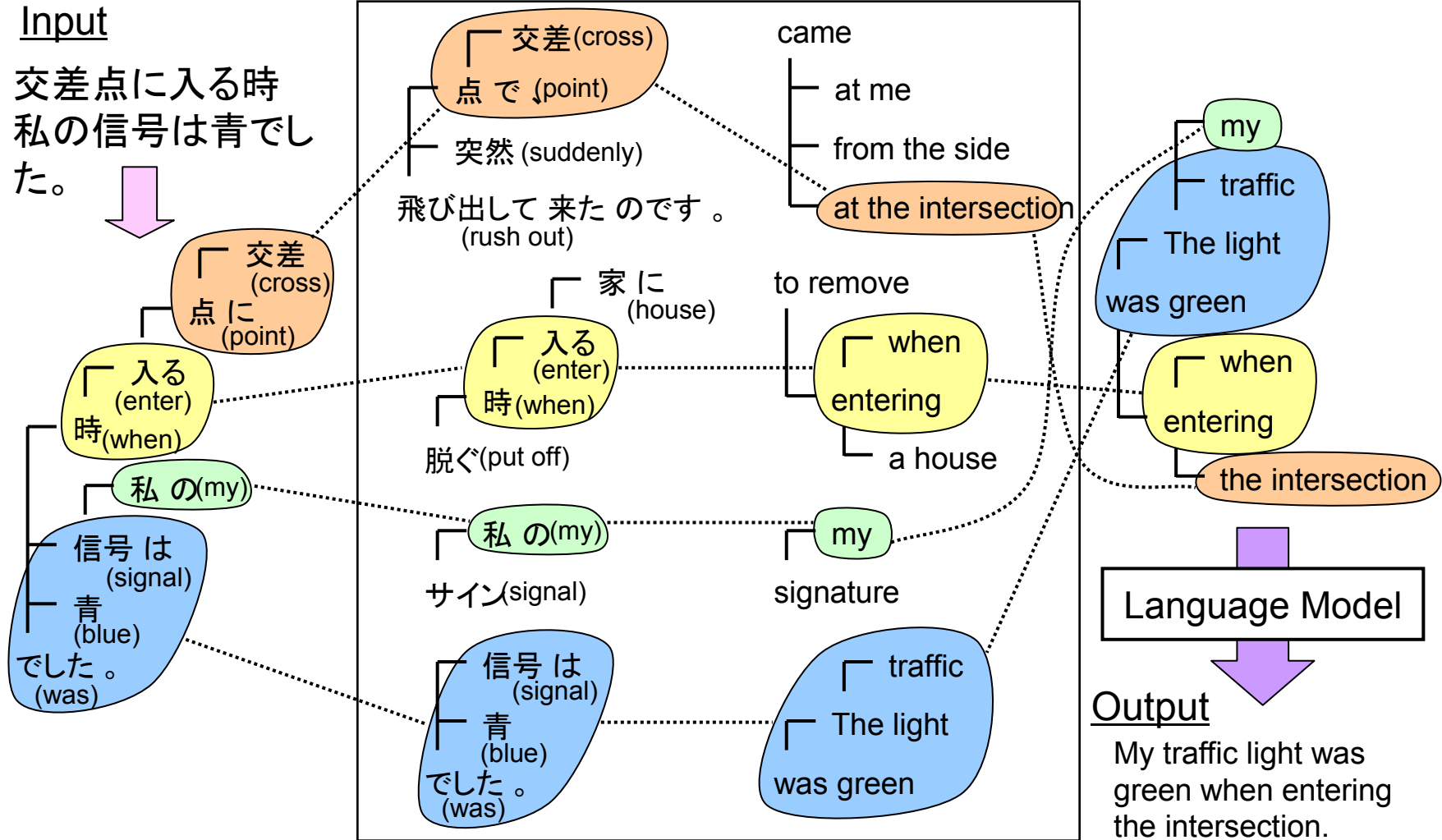
- Why EBMT?
- **Description of Kyoto-U EBMT System**
- Japanese Particular Processing
  - Pronoun Estimation
  - Japanese Flexible Matching
- Result and Discussion
- Conclusion and Future Work

# Kyoto-U System Overview

## Translation Examples

### Input

交差点に入る時  
私の信号は青で  
した。



### Output

My traffic light was green when entering the intersection.

## **Structure-based Alignment**

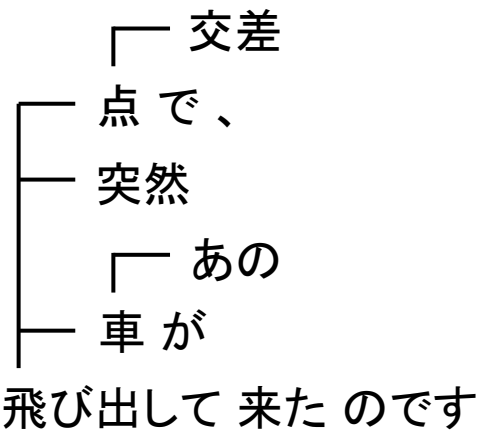
- **Step1: Dependency structure transformation**
- **Step2: Word/phrase correspondences detection**
- **Step3: Correspondences disambiguation**
- **Step4: Handling remaining words**
- **Step5: Registration to database**

Step1

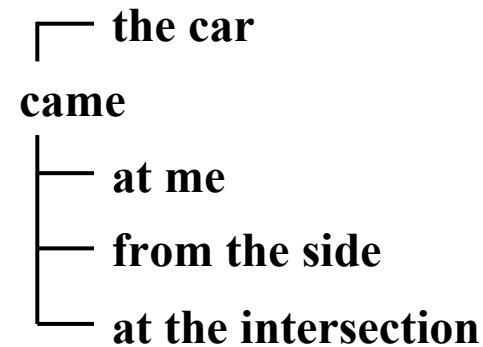
# Dependency Structure Transformation

- J: JUMAN/KNP
- E: Charniak's nlparsner → Dependency tree

J: 交差点で、突然あの車が  
飛び出して来たのです。



E: The car came at me from  
the side at the intersection.



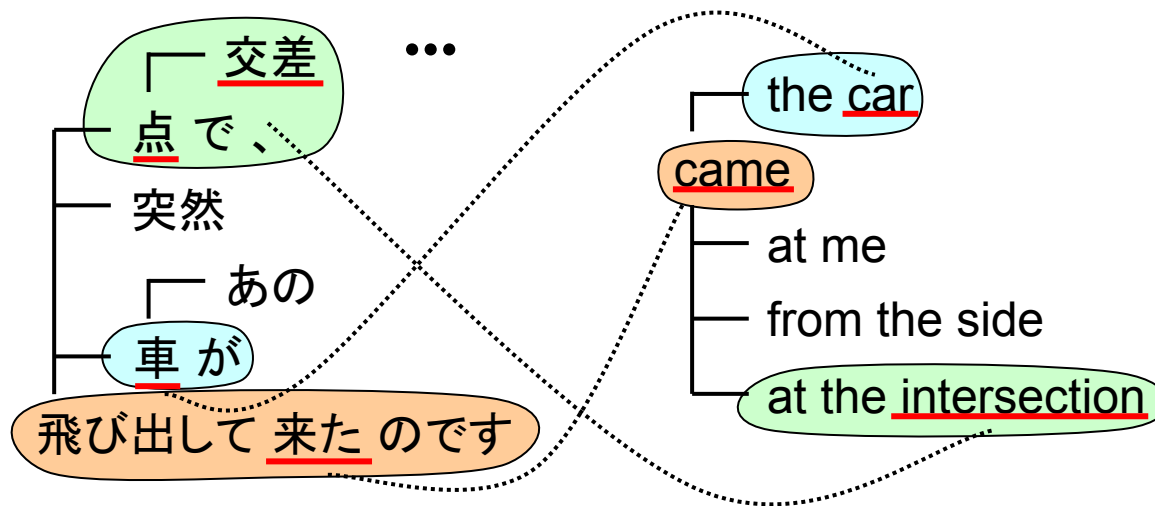


Step2

## Word Correspondence Detection

- KENKYUSYA J-E, E-J dictionaries (300K entries)
- Transliteration (person/place names, Katakana words)

Ex) 新宿 → shinjuku ⇔ shinjuku (similarity:1.0)  
sinjuku  
synjucu



Step3

## Correspondence Disambiguation

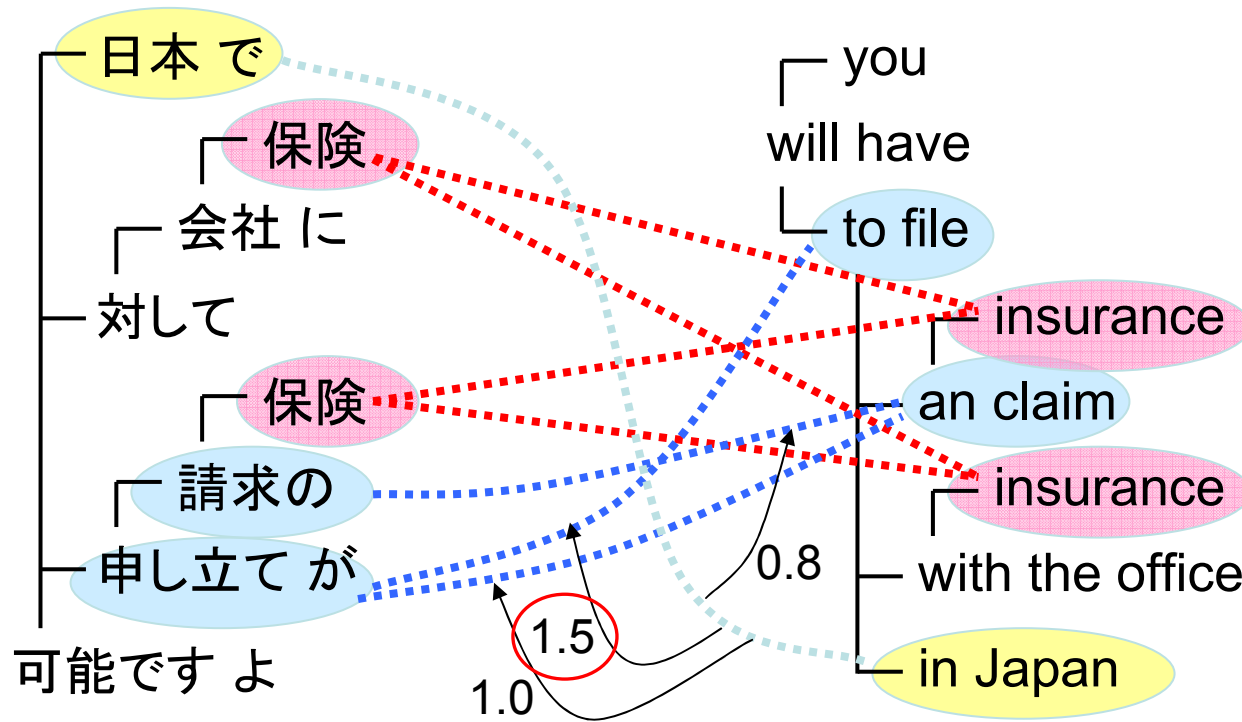
- Calculate correspondence score based on unambiguous alignment
- Select correspondence with higher score

$$\text{Score} = \sum_{\text{Unamb. Matches}} \frac{1}{\text{dist}_J} + \frac{1}{\text{dist}_E}$$

$\text{dist}_{J/E}$  = Distance to unambiguous correspondence  
in Japanese/English tree

Step3

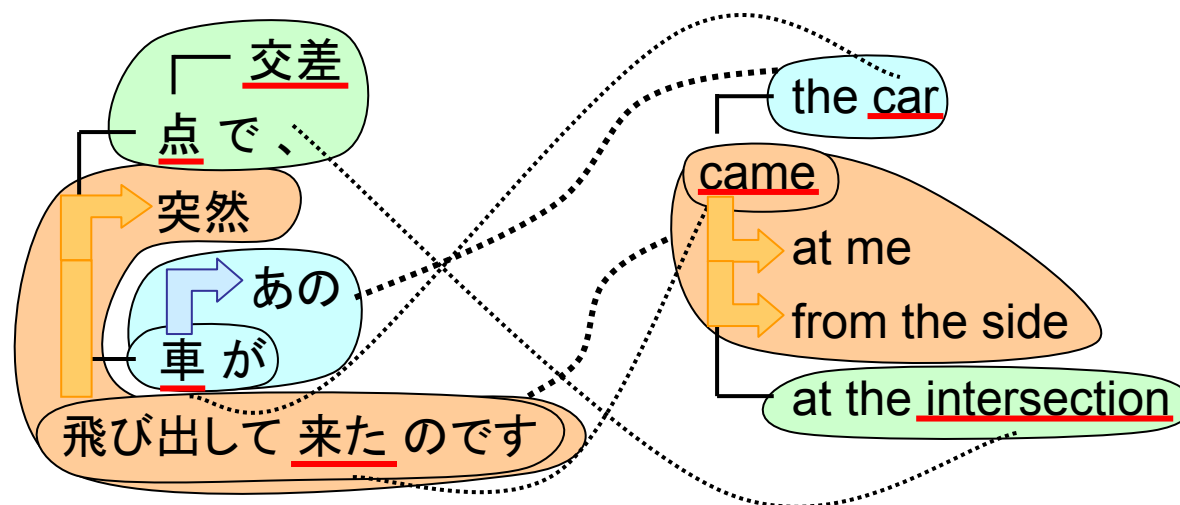
## Correspondence Disambiguation (cont.)



Step4

## Handling Remaining Words

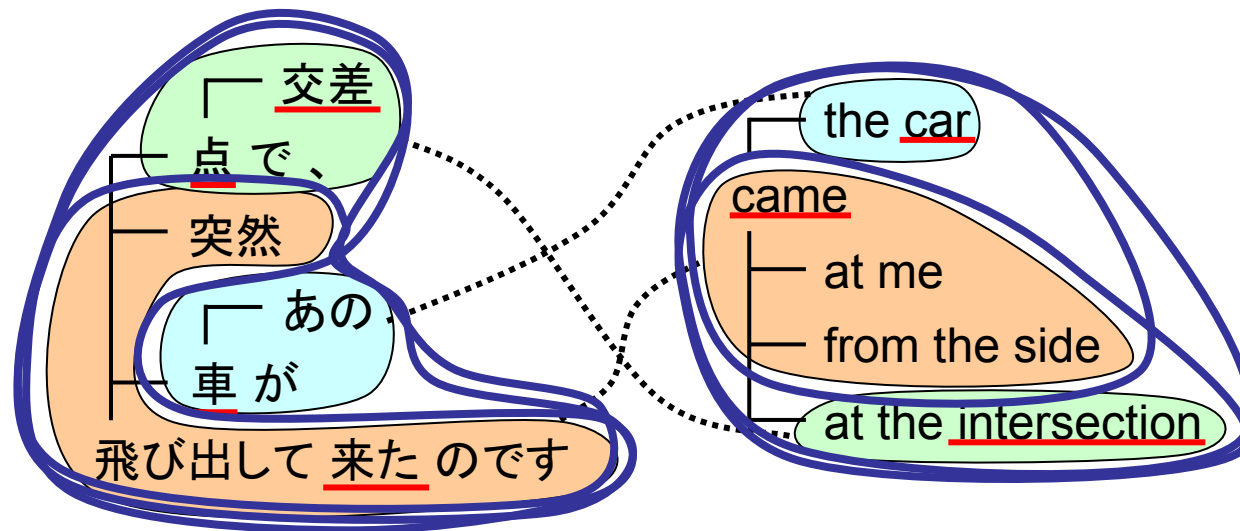
- Align root nodes when remained
- Merge Base NP nodes
- Merge into ancestor nodes



Step5

## Registration to Database

- Register each correspondence
- Register a couple of correspondences

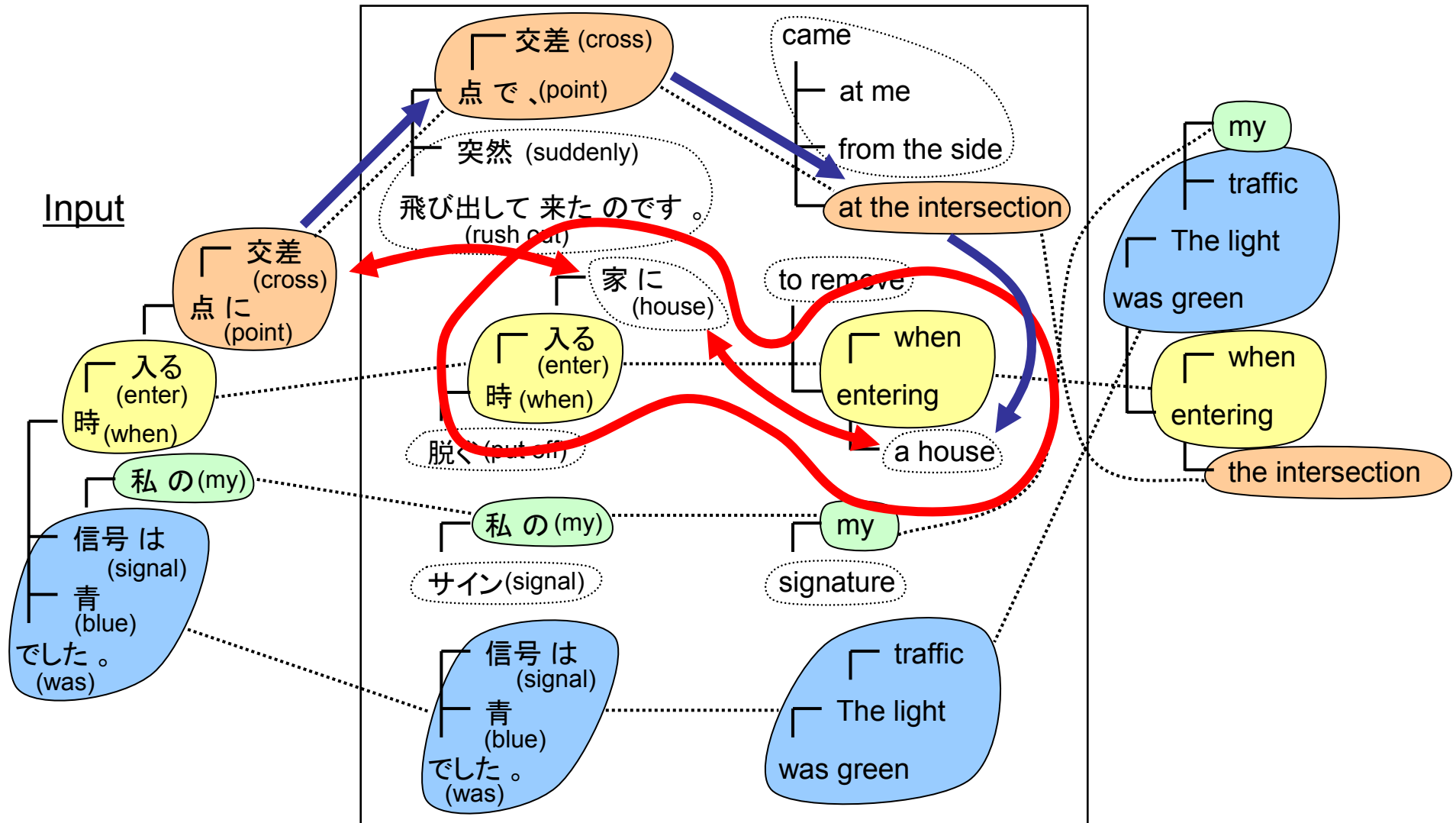


# **Translation**

- **Translation example (TE) retrieval**
  - **for all the sub-trees in the input**
- **TE selection**
  - **prefer to large size example**
- **TE combination**
  - **greedily form the root node**

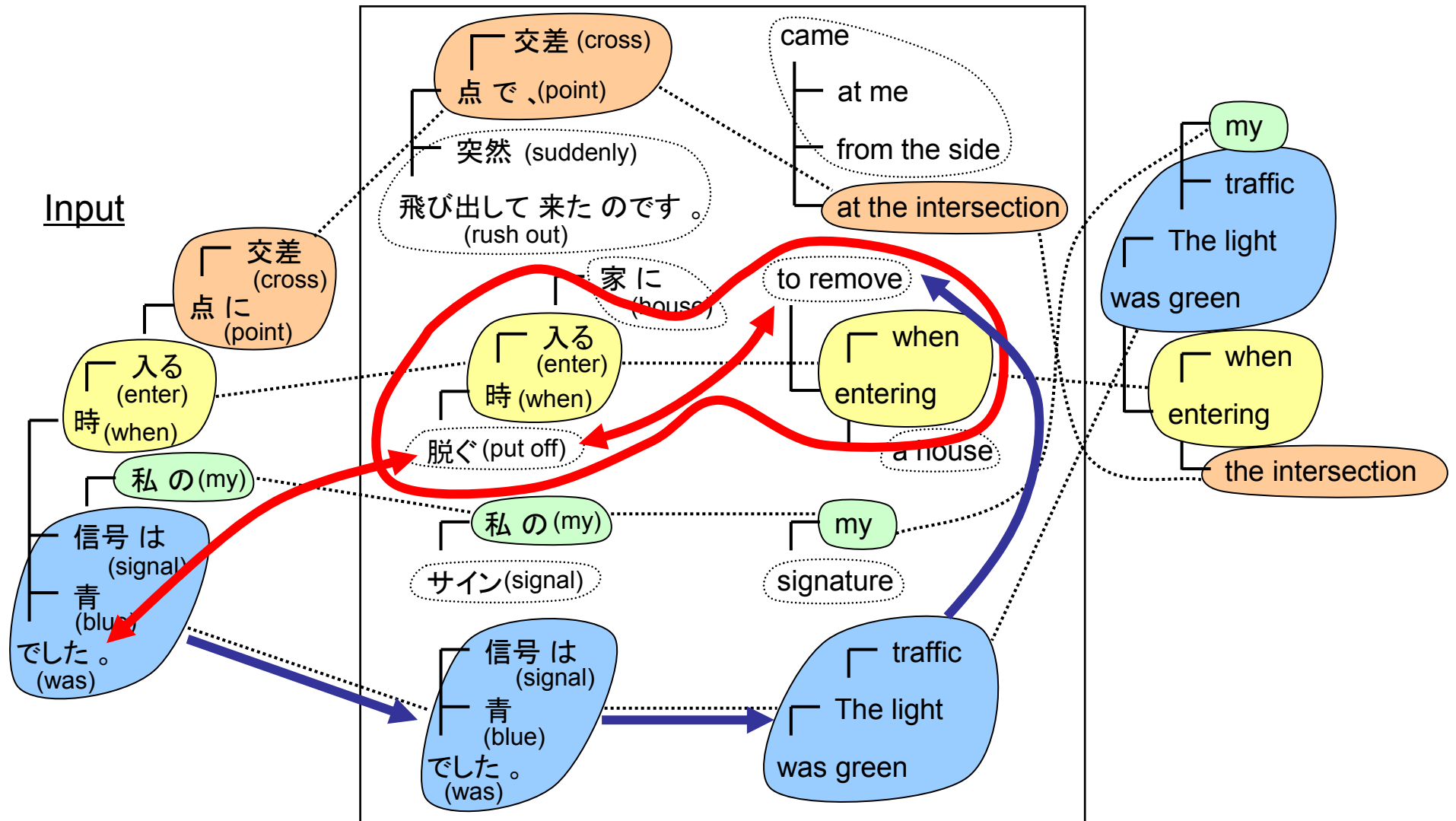
# Combination Example

## Translation Examples



# Combination Example (cont.)

## Translation Examples





# Outline

- Why EBMT?
- Description of Kyoto-U EBMT System
- **Japanese Particular Processing**
  - **Pronoun Estimation**
  - **Japanese Flexible Matching**
- Result and Discussion
- Conclusion and Future Work

# Pronoun Estimation

## ➤ Pronouns are often omitted in Japanese sentences

### ✓ Omitted in TE:

- TE

胃が痛いのです → I've a stomachache

- Input

私は胃が痛いのです → I I've a stomachache ✗

### ✓ Omitted in Input

- TE

これを日本に送ってください → Will you mail **this** to Japan?

- Input:

日本へ送ってください → Will you mail to Japan? ✗



## Pronoun Estimation (cont.)

### ➤ Estimate omitted pronoun by modality and subject case

#### ✓ Omitted in TE:

- TE

(私は)胃が痛いのです → I've a stomachache

- Input

私は胃が痛いのです → I've a stomachache ○

#### ✓ Omitted in Input

- TE

これを日本に送ってください → Will you mail **this** to Japan?

- Input:

(これを)日本へ送ってください → Will you mail this to Japan? ○

# Various Expressions in Japanese

## ➤ Synonymous Relation

- Hiragana/Katakana/Kanji variations

りんご = リンゴ = 林檎 (apple)      **Morphological**

- Variations of Katakana expressions      **Analyzer**

コンピュータ = コンピューター (computer)

- Synonymous words

登山 = 山登り (climbing mountain vs mountain climbing)

- Synonymous phrases

最寄りの = 一番近い  
(nearest)      (most) (near)

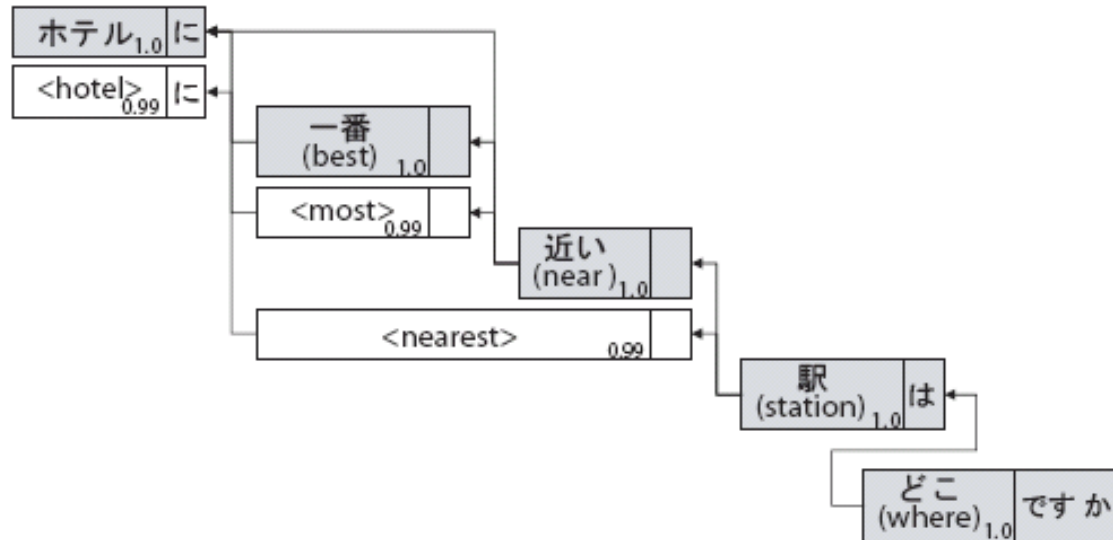
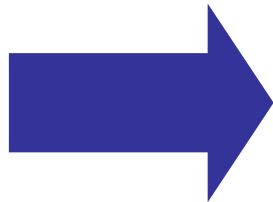
**Automatically  
Acquired from  
Japanese  
Dictionaries**

## ➤ Hypernym-Hyponym Relation

- 災難 ← 災害 ← 地震(earthquake)、台風(typhoon)  
(disaster)

# Japanese Flexible Matching

- ホテルに 一番 近い 駅 は どこですか  
(hotel to best near station TOP where is)
- = ホテルに 最も 近い 駅 は どこですか  
(hotel to most near station TOP where is)
- = ホテルの 最寄り の 駅 は どこですか  
(hotel to nearest of station TOP where is)
- = 旅館に *ichiban* 近い 駅 は どこですか  
(hotel to best near station TOP where is)
- ≡ ホテルに 近い 駅 は どこですか  
(hotel to near station TOP where is)
- ≡ ...



# IWSLT06 Evaluation Results

- Open data track (JE)
- Correct recognition translation & ASR output translation

		BLEU	NIST
Correct recognition	Dev1	0.5087	9.6803
	Dev2	0.4881	9.4918
	Dev3	0.4468	9.1883
	Dev4	0.1921	5.7880
	Test	0.1655 (8 <sup>th</sup> /14)	5.4325 (8 <sup>th</sup> /14)
ASR output	Dev4	0.1590	5.0107
	Test	0.1418 (9 <sup>th</sup> /14)	4.8804 (10 <sup>th</sup> /14)

# Results Discussion

- **Punctuation insertion failure caused parsing error**
- **Dictionary robustness affected alignment accuracy**
- **TE selection criterion failed when choosing among ‘almost equal’ examples**
  - e.g. Input: “買います” (buy a ticket)  
TE: “買いません” (**not** buy a ticket)

## **Conclusion and Future Work**

- **We not only aim at the development of MT, but also tackle this task from the viewpoint of structural NLP.**
- **Implement statistical method on alignment**
- **Improve parsing accuracies (both J and E)**
- **Improve Japanese flexible matching method**
- **J-C and C-J MT Project with NICT**