Example-based Machine Translation based on Deeper NLP

Toshiaki Nakazawa[1], Kun Yu[1], Sadao Kurohashi[2]

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

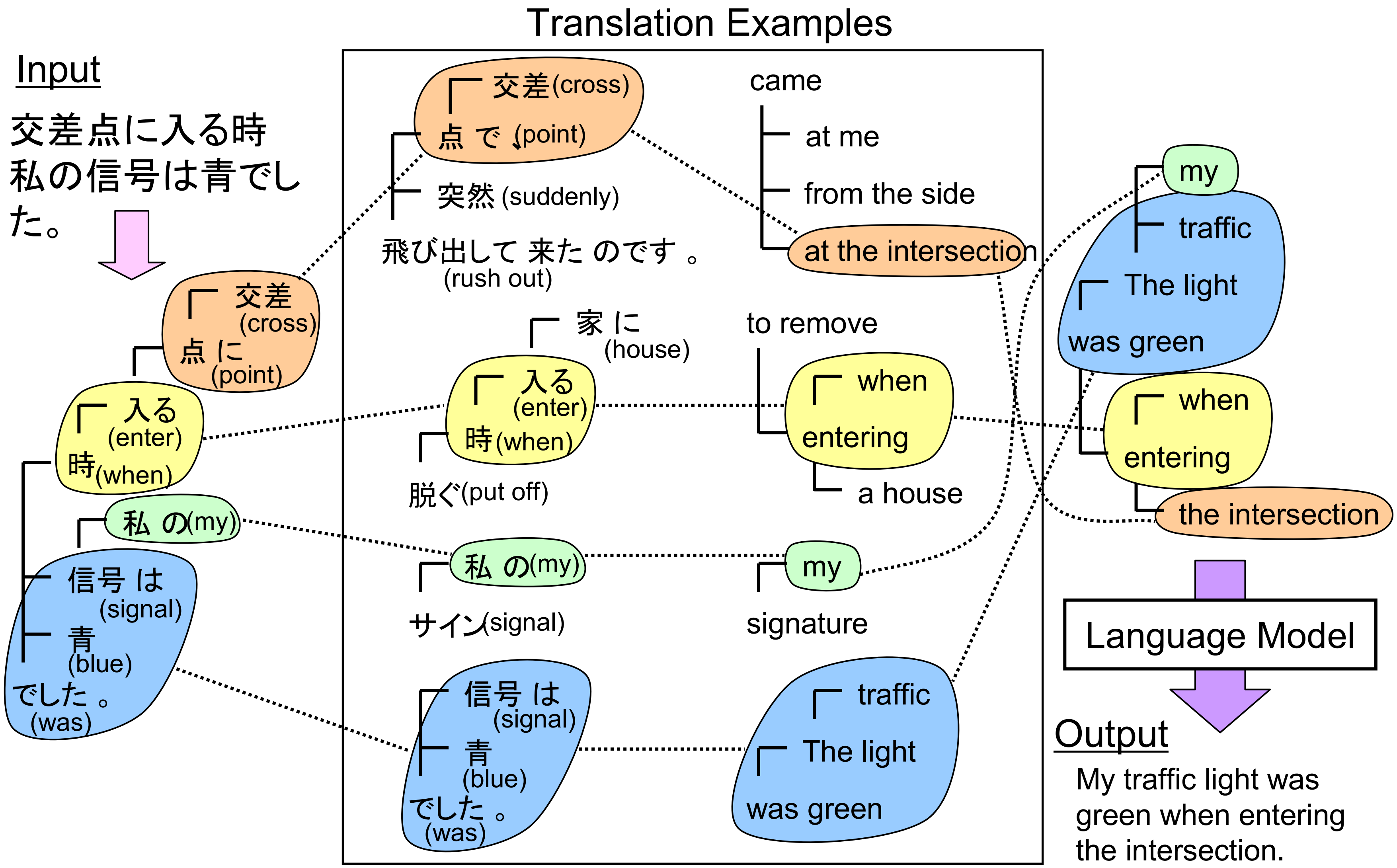
Why EBMT?

- 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



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

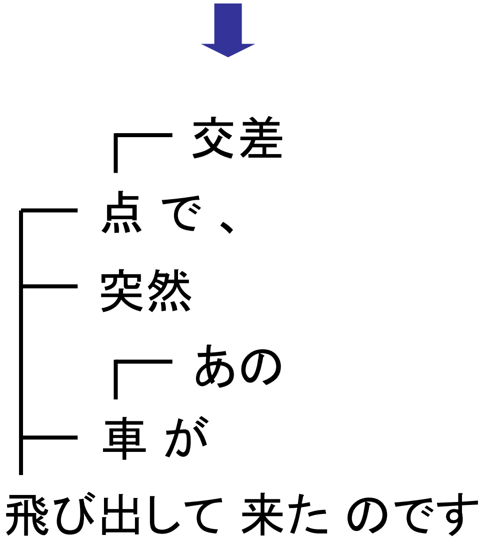
Step1

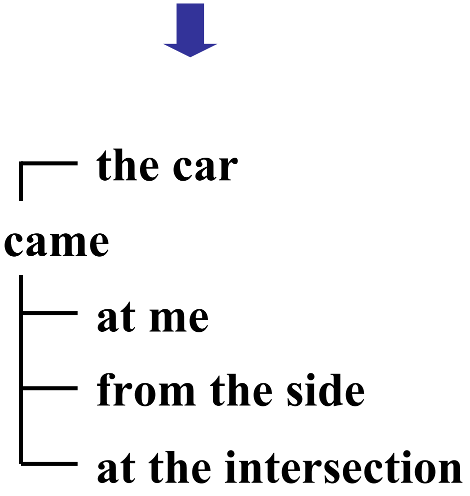
Dependency Structure Transformation

- J: JUMAN/KNP
- E: Charniak’s nlparser → 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.


交差
点 で 、
突然
あの
車 が
飛び出して 来た のです

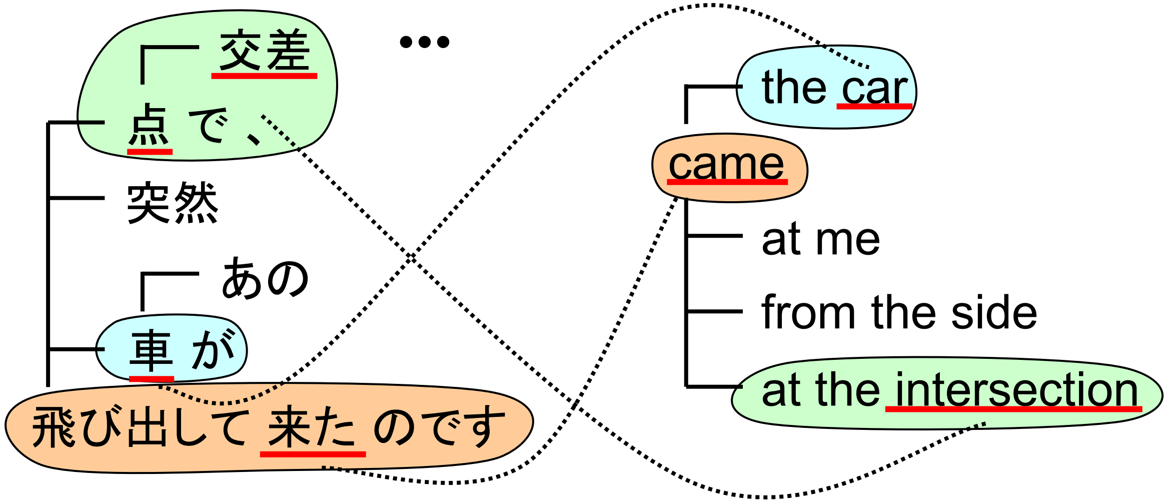

the car
came
at me
from the side
at the intersection

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

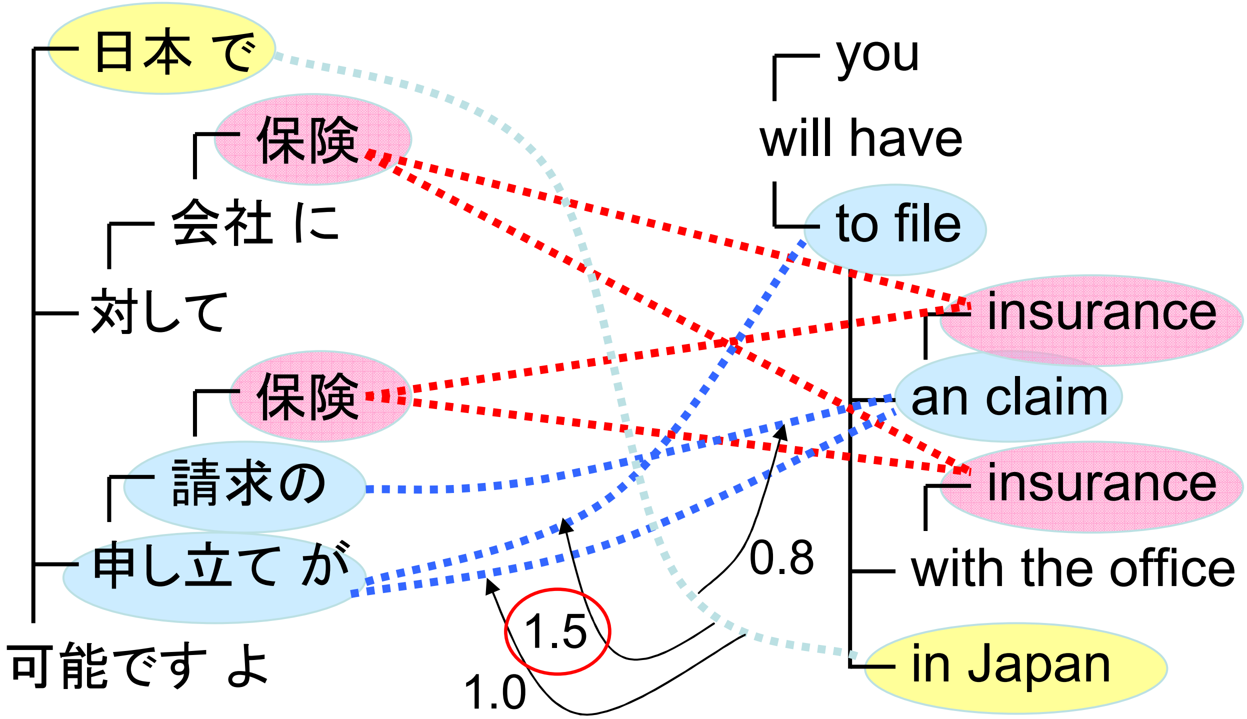
- **Calculate correspondence score based on unambiguous alignment**
- **Select correspondence with higher score**

$$\mathrm{Score} = \sum_{Unamb.Matches} \frac{1}{dist_J} + \frac{1}{dist_E}$$

$dist_{J/E}$ = Distance to unambiguous correspondence
in Japanese/English tree

Step3

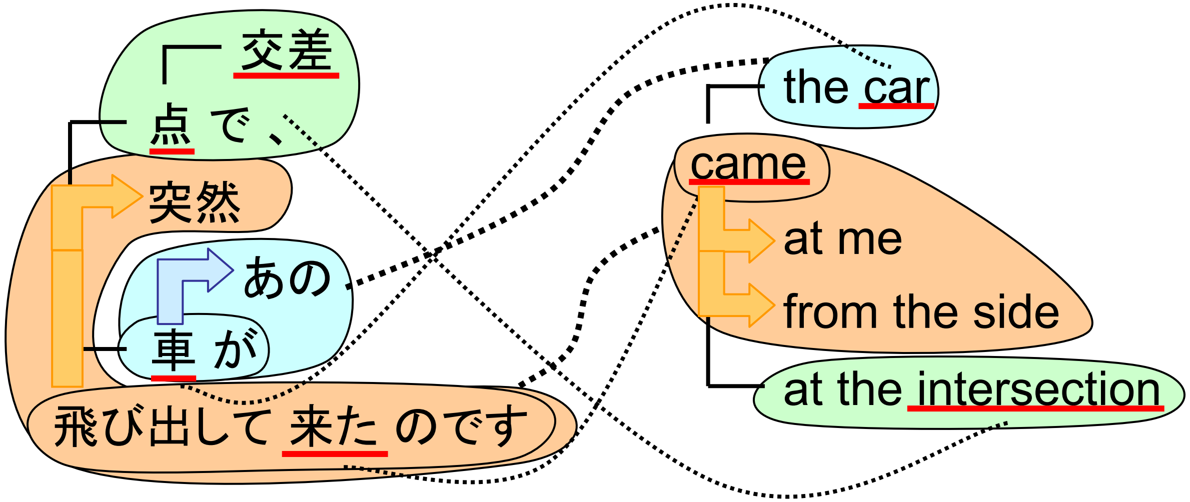
Correspondence Disambiguation (cont.)



Step4

Handling Remaining Words

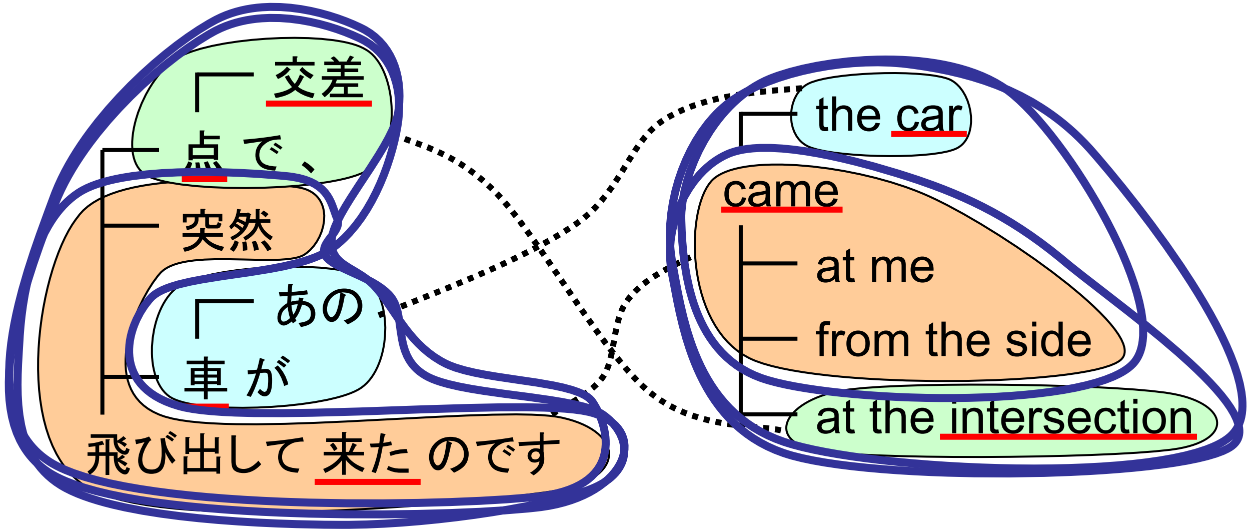
- Align root nodes when remained
- Merge Base NP nodes
- Merge into ancestor nodes



Step5

Registration to Database

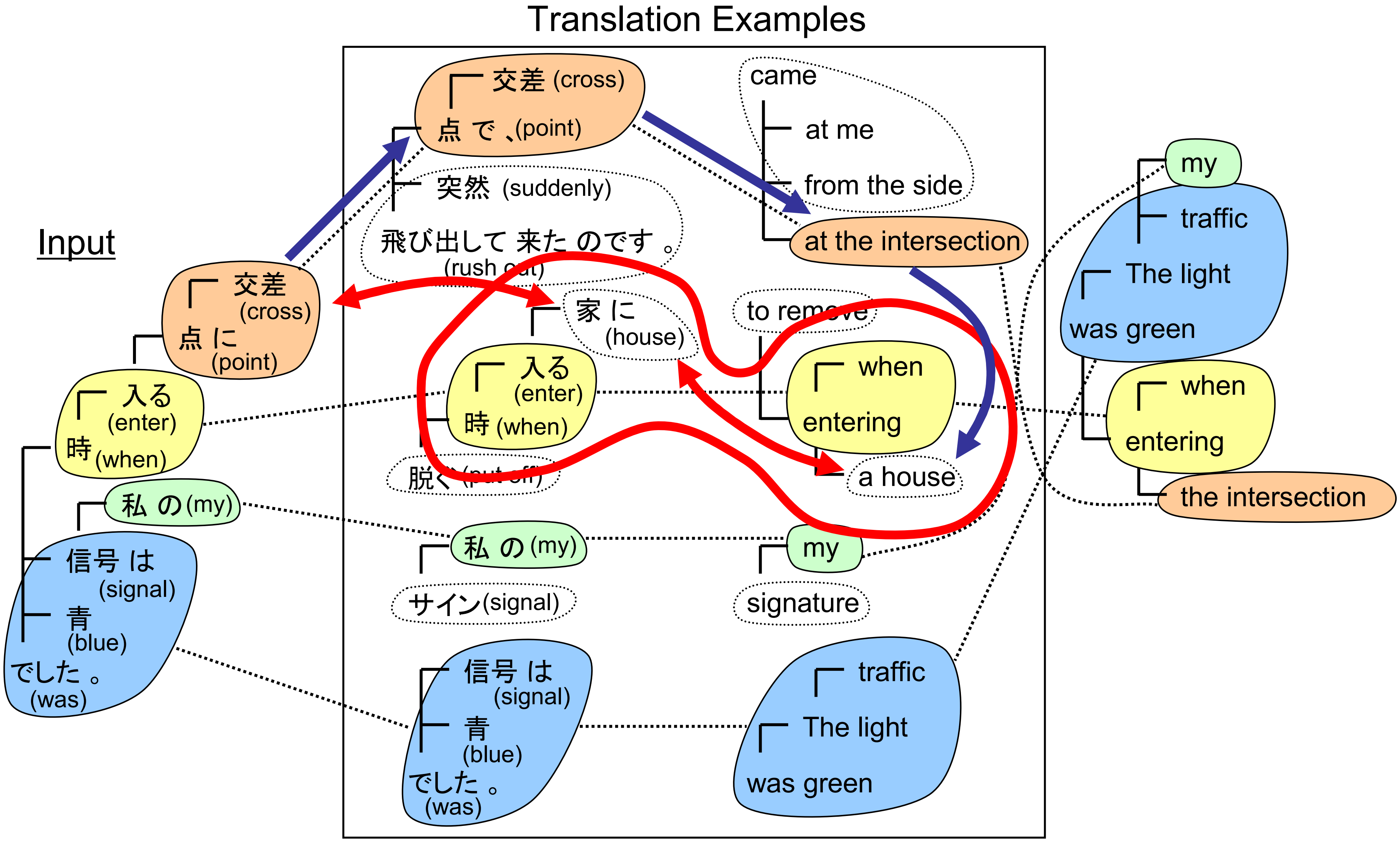
- Register each correspondence
- Register a couple of correspondences



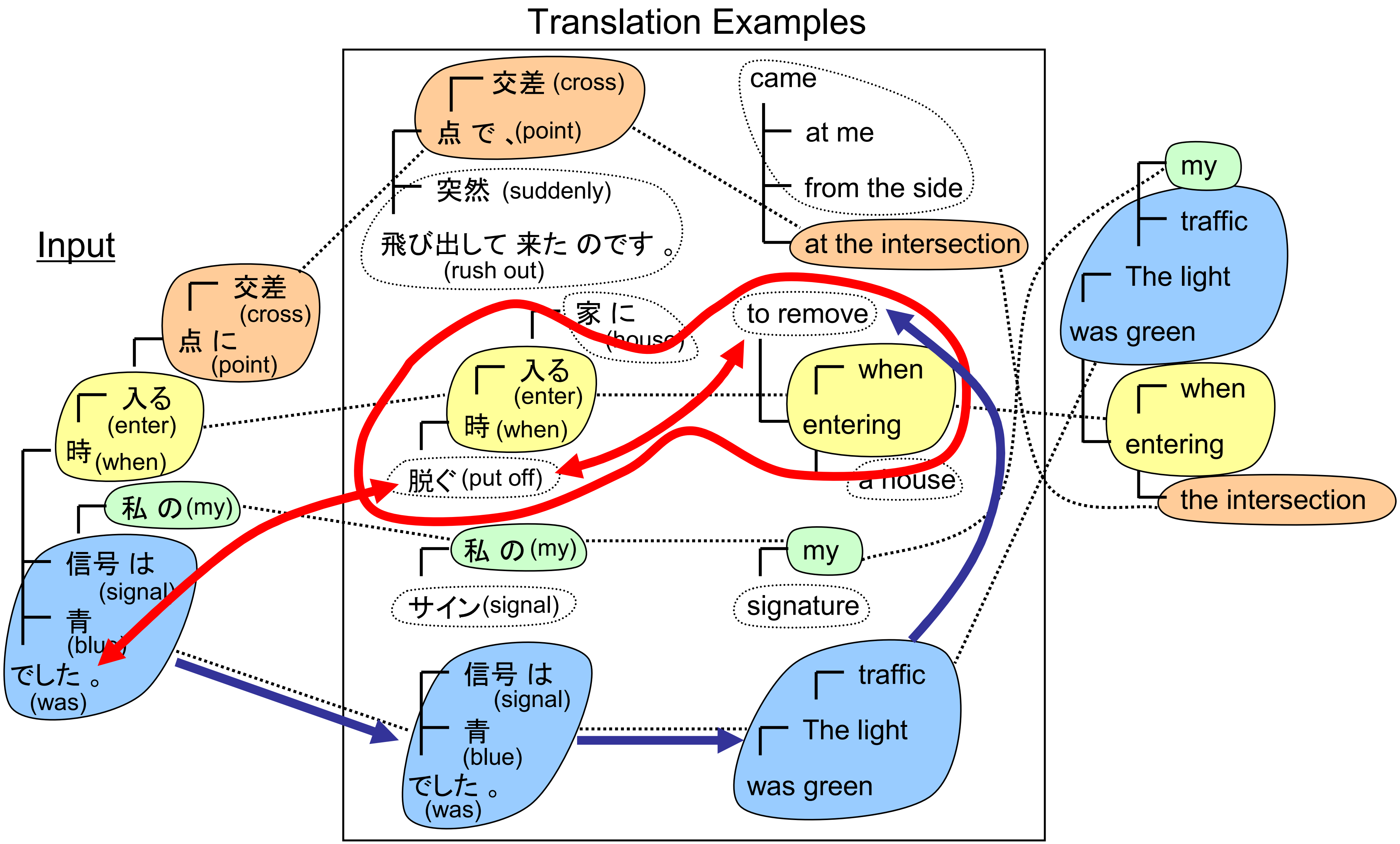
Translation

- 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



Combination Example (cont.)



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)
  - Variations of Katakana expressions

    コンピュータ ＝ コンピューター (computer)

Morphological Analyzer

  - Synonymous words

    登山 = 山登り (climbing mountain vs mountain climgbing)
  - Synonymous phrases

    最寄りの = 一番近い

    (nearest) (most) (near)

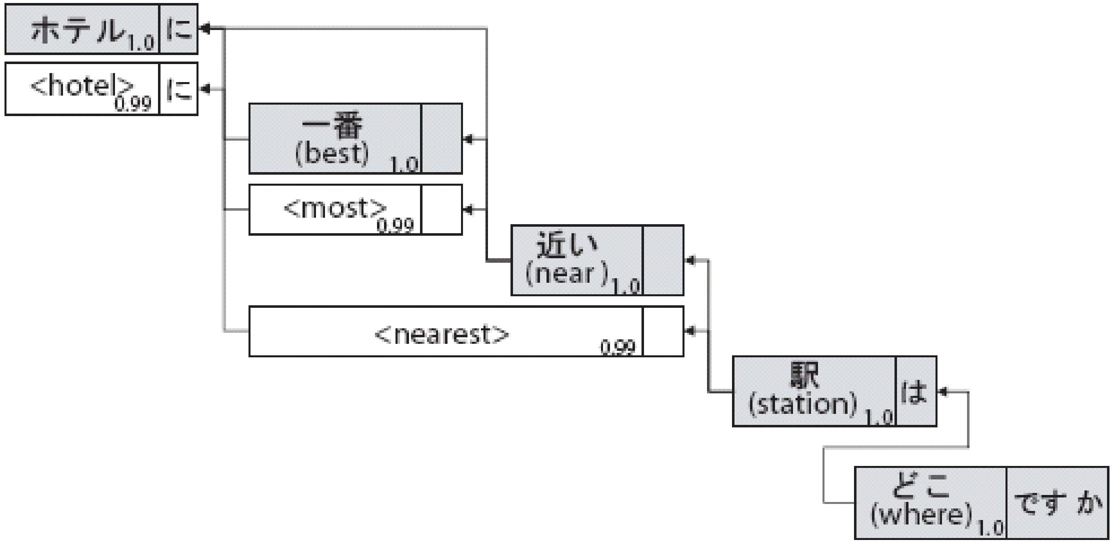
- Hypernym-Hyponym Relation
  - 災難 ← 災害 ← 地震(earthquake)、台風(typhoon)

    (disaster)

Automatically Acquired from Japanese Dictionaries

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 (8th/14)	5.4325 (8th/14)
ASR output	Dev4	0.1590	5.0107
	Test	0.1418 (9th/14)	4.8804 (10th/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