Example-based Machine Translation based on Deeper NLP

Toshiaki Nakazawa¹, Kun Yu¹, Sadao Kurohashi²

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
The light was green when entering the intersection.
Structure-based Alignment

- Step 1: Dependency structure transformation
- Step 2: Word/phrase correspondences detection
- Step 3: Correspondences disambiguation
- Step 4: Handling remaining words
- Step 5: Registration to database
Dependency Structure Transformation

- J: JUMAN/KNP
- E: Charniak’s nlparsen → 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.
Step 2

**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)
sunjuku synjucu

The car came at me from the side at the intersection.
Step 3

**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
Correspondence Disambiguation (cont.)

1. You will have to file an insurance claim with the office in Japan.

Step 3
Step 4

Handling Remaining Words

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

Translated Text:

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
Step 5

**Registration to Database**

- Register each correspondence
- Register a couple of correspondences

交差点で、突然あの車が飛び出してきたのです

the car came at me from the side at the intersection
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

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.

Translation Examples

The light was green when entering the intersection.
The light was green when entering the intersection. My traffic light was blue when I entered the house suddenly at the intersection.
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

- Hiragana/Katakana/Kanji variations
  りんご = リンゴ = 林檎 (apple)  Morphological Analyzer

- Variations of Katakana expressions
  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

<table>
<thead>
<tr>
<th>Correct recognition</th>
<th>BLEU</th>
<th>NIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dev1</td>
<td>0.5087</td>
<td>9.6803</td>
</tr>
<tr>
<td>Dev2</td>
<td>0.4881</td>
<td>9.4918</td>
</tr>
<tr>
<td>Dev3</td>
<td>0.4468</td>
<td>9.1883</td>
</tr>
<tr>
<td>Dev4</td>
<td>0.1921</td>
<td>5.7880</td>
</tr>
<tr>
<td>Test</td>
<td>0.1655 (8th/14)</td>
<td>5.4325 (8th/14)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ASR output</th>
<th>BLEU</th>
<th>NIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dev4</td>
<td>0.1590</td>
<td>5.0107</td>
</tr>
<tr>
<td>Test</td>
<td>0.1418 (9th/14)</td>
<td>4.8804 (10th/14)</td>
</tr>
</tbody>
</table>
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