Example-based Machine Translation
Pursuing Fully Structural NLP

Sadao Kurohashi, Toshiaki Nakazawa, Kauffmann Alexis, Daisuke Kawahara
University of Tokyo
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.
Overview of UTokyo System

Translation Examples

**Input**

交差点に入ると私の信号は青でした。

**Output**

My traffic light was green when entering the intersection.
I. Background
II. Alignment of Parallel Sentences
III. Translation
IV. Beyond Simple EBMT
V. IWSLT Results and Discussion
VI. Conclusion
Common Feature

- Use bilingual corpus, or translation examples for the translation of new inputs.
- Exploit translation knowledge implicitly embedded in bilingual corpus.
- Make MT system maintenance and improvement much easier compared with Rule-based MT.
<table>
<thead>
<tr>
<th>SMT</th>
<th>EMBT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Problem setting</strong></td>
<td></td>
</tr>
<tr>
<td>- Only bilingual corpus</td>
<td>- Any resources (bilingual corpus are not necessarily huge)</td>
</tr>
<tr>
<td><strong>Methodology</strong></td>
<td></td>
</tr>
<tr>
<td>- Combine words/phrases with high probability</td>
<td>- Try to use larger translation examples (→ syntactic information)</td>
</tr>
</tbody>
</table>
Why EBMT?

- Pursuing structural NLP
  - Improvement of basic analyses leads to improvement of MT
  - Feedback from application (MT) can be expected

- EMBT setting is suitable in many cases
  - Not a large corpus, but similar examples in relatively close domain
    - Translation of manuals using the old version manuals’ translation
    - Patent translation using related patents’ translation
    - Translation of an article using the already translated sentences step by step
I. Background
II. Alignment of Parallel Sentences
III. Translation
IV. Beyond Simple EBMT
V. IWSLT Results and Discussion
VI. Conclusion
1. Transformation into dependency structure

J: JUMAN/KNP
E: Charniak’s nlpars → Dependency tree
1. Transformation into dependency structure
2. Detection of word(s) correspondences
   - EIJIRO (J-E dictionary): 0.9M entries
   - Transliteration detection
     ローズワイン → rosuwain ⇔ rose wine (similarity:0.78)
     新宿 → shinjuku ⇔ shinjuku (similarity:1.0)
1. Transformation into dependency structure
2. Detection of word(s) correspondences
3. Disambiguation of correspondences
In the 20,000 J-E training data, ambiguous correspondences are only 4.8%.

\[ C_{unamb} \rightarrow C_{amb} : 1/(Distance \text{ in } J \text{ tree}) + 1/(Distance \text{ in } E \text{ tree}) \]
1. Transformation into dependency structure
2. Detection of word(s) correspondences
3. Disambiguation of correspondences
4. Handling of remaining phrases
   1. The root nodes are aligned, if remaining
   2. Expansion in base NP nodes
   3. Expansion downwards

Intersection where suddenly that car came at me from the side at the intersection
1. Transformation into dependency structure
2. Detection of word(s) correspondences
3. Disambiguation of correspondences
4. Handling of remaining phrases
5. Registration to translation example database
I. Background
II. Alignment of Parallel Sentences
III. Translation
IV. Beyond Simple EBMT
V. IWSLT Results and Discussion
VI. Conclusion
The light was green when entering the intersection.

Language & Knowledge Engineering Lab

Translation Examples

交差点に入る時私の信号は青でした。

Input

交差点 (cross)
点に(point)
入れる(enter)
時(when)
私の(my)
信号(信号)
青(blue)
でした(was)

突然(suddenly)
飛び出して来たのです。(rush out)

家に(house)
入れる(enter)
時(when)
脱ぐ(put off)
私の(my)
信号(信号)
青(blue)
でした(was)

Output

My traffic light was green when entering the intersection.
1. Retrieval of translation examples
   For all the sub-trees in the input

1. Selection of translation examples
   The criterion is based on the size of translation example (the number of matching nodes with the input), plus the similarities of the neighboring outside nodes. ([Aramaki et al. 05] proposed a selection criterion based on translation probability.)

1. Combination of translation examples
Combining TEs using Bond Nodes

Translation Examples

Input

交差 (cross) 交差 (cross)
点 で 、(point)

突然 (suddenly) 飛び出して 来た のです。 (rush out)

入る (enter) 家 に (house)
時 (when) 脱ぐ <cut off>

信号 は (signal) 私 の (my)
青 (blue) でした 。 (was)

The light was green
交通 (traffic)
私の サイン (signature)

my traffic came at me from the side
at the intersection
to remove
at entering
when entering
a house

交差点 に (point)

Translation Examples

Input

交差 (cross) 交差 (cross)
点 で 、(point)

突然 (suddenly) 飛び出して 来た のです。 (rush out)

入る (enter) 家 に (house)
時 (when) 脱ぐ <cut off>

信号 は (signal) 私 の (my)
青 (blue) でした 。 (was)

The light was green
交通 (traffic)
私の サイン (signature)

my traffic came at me from the side
at the intersection
to remove
at entering
when entering
a house

交差点 に (point)

Translation Examples

Input

交差 (cross) 交差 (cross)
点 で 、(point)

突然 (suddenly) 飛び出して 来た のです。 (rush out)

入る (enter) 家 に (house)
時 (when) 脱ぐ <cut off>

信号 は (signal) 私 の (my)
青 (blue) でした 。 (was)

The light was green
交通 (traffic)
私の サイン (signature)

my traffic came at me from the side
at the intersection
to remove
at entering
when entering
a house

交差点 に (point)
Combining TEs using Bond Nodes

Translation Examples

Input

The light was green when entering the intersection, my traffic came at me from the side.

Translation

The light was green when entering the intersection, my traffic came at me from the side.

Input

The light was green when entering the intersection, my traffic came at me from the side.

Translation

The light was green when entering the intersection, my traffic came at me from the side.

Input

The light was green when entering the intersection, my traffic came at me from the side.

Translation

The light was green when entering the intersection, my traffic came at me from the side.

Input

The light was green when entering the intersection, my traffic came at me from the side.

Translation

The light was green when entering the intersection, my traffic came at me from the side.
Outline

I. Background
II. Alignment of Parallel Sentences
III. Translation
IV. Beyond Simple EBMT
V. IWSLT Results and Discussion
VI. Conclusion
- **Cardinal**: 124 → one hundred twenty four
- **Ordinal (e.g., day)**: 2 日 → second
- **Two-figure (e.g., room #, year)**: 124 → one twenty four
- **One-figure (e.g., flight #, phone #)**: 124 → one two four
- **Non-numeral (e.g., month)**: 8 月 → August
Pronoun Omission

- TE:
  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

- 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 this to Japan?
I. Background
II. Alignment of Parallel Sentences
III. Translation
IV. Beyond Simple EBMT
V. IWSLT Results and Discussion
VI. Conclusion
### Evaluation Results

Supplied 20,000 JE data, Parser, Bilingual dictionary  
*(Supplied + tools ; Unrestricted)*

<table>
<thead>
<tr>
<th></th>
<th>BLUE</th>
<th>NIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dev 1</td>
<td>0.424</td>
<td>8.57</td>
</tr>
<tr>
<td>Dev 2</td>
<td>0.405</td>
<td>8.50</td>
</tr>
<tr>
<td>IWSLT05 Manual</td>
<td>0.372</td>
<td>7.85</td>
</tr>
<tr>
<td></td>
<td><em>(4^{th}/7; 2^{nd}/3)</em></td>
<td><em>(3^{rd}/7; 2^{nd}/3)</em></td>
</tr>
<tr>
<td>IWSLT05 ASR</td>
<td>0.336</td>
<td>7.42</td>
</tr>
</tbody>
</table>
Translation of a test sentence
- 7.5 words/3.2 phrases
- 1.8 TEs of the size of 1.5 phrases + 0.5 translation from dic.

Parsing accuracy (100 sent.)
- J: 94%, E: 77% (sentence level)

Alignment precision (100 sent.)
- Word(s) alignment by bilingual dictionary: 92.4%
- Phrase alignment: 79.1%
  ⇔ Giza++ one way alignment: 64.2%

“Is the current parsing technology useful and accurate enough for MT?”
Conclusion

- We not only aim at the development of MT, but also tackle this task from the viewpoint of structural NLP.

- Future work
  - Improve paring accuracies of both languages complementary
  - Flexible matching in monolingual texts
  - Anaphora resolution
  - J–C and C–J MT Project with NICT