Word Selection for EBMT based on Monolingual Similarity and Translation Confidence

Eiji Aramaki * **
Hideki Kashioka **

Sadao Kurohashi * **
Hideki Tanaka **

* University of Tokyo
** ATR Spoken Language Translation Research Laboratories
An EBMT system requires a Translation Memory.
EBMT Framework

If highly parallel corpus is available
construction of TM is relatively easy
EBMT Framework

Content Aligned Corpus ➔ Translation Memory ➔ EBMT system

Input sentence ➔ Output sentence

Most available corpora are share the same content. (e.g. newspapers, broadcast-news)
Outline

1: NHK News Corpus

2: How to build a Translation Memory (TM)

3: How to use the TM

4: Experiments

5: Conclusion
Ambassadors and diplomats from 37 countries took part in a rice planting festival on Sunday in small paddies on steep hillsides in Wajima, central Japan.

About one-thousand people gathered at the hill, where some two-thousand 100 miniature paddies, called Senmaida, stretch toward the Sea of Japan.

The event was organized by the private Foundation for Global Peace and Environment.

The rice seedlings are grown from grain harvested by the Emperor at the Imperial Palace in Tokyo last autumn.

Barefoot participants waded into the paddies to plant the seedlings by hand while singing a local folk song about the practice of rice planting.
Ambassadors and diplomats from 37 countries took part in a rice planting festival on Sunday in small paddies on steep hillsides in Wajima, central Japan.

About one-thousand people gathered at the hill, where some two-thousand 100 miniature paddies, called Senmaida, stretch toward the Sea of Japan.

The event was organized by the private Foundation for Global Peace and Environment.

The rice seedlings are grown from grain harvested by the Emperor at the Imperial Palace in Tokyo last autumn.

Barefoot participants waded into the paddies to plant the seedlings by hand while singing a local folk song about the practice of rice planting.

The Imperial Palace "in Tokyo"

Schedule of next year event
1: NHK News Corpus

2: How to build a Translation Memory (TM)

3: How to select Translation Examples from TM

4: Experiments

5: Conclusion
**Sentence Alignment**

- DP matching method using 5 translation dictionaries (200,000 entries in total)
- Extract 1-to-1 sentence-pairs
  - 1:1-sentence-pairs have higher accuracy than the others

<table>
<thead>
<tr>
<th>Japanese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>田植えフェスティバル石川県輪島市で外国の大使や一般の参加者など千人あまりが急な斜面の棚田で田植えを体験する催しが行われました。</td>
<td>Ambassadors and diplomats from 37 countries took part in a rice planting festival on Sunday in small paddies on steep hillsides in Wajima, central Japan.</td>
</tr>
<tr>
<td>輪島市白米町には(しろよねまち)千枚田と呼ばれる(せんまいだ)大小二千枚の棚田が急な斜面から海に向かって拡がっています。</td>
<td>About one-thousand people gathered at the hill, where some two-thousand 100 miniature paddies, called Senmaida, stretch toward the Sea of Japan.</td>
</tr>
<tr>
<td>田植え体験は農作業を通じて米作りの意義などを考えているフェスティバル経過平和財団の呼び掛けで開かれたもので、海外三十国の大使や書記官、それに一般の参加者ら合わせておよそ千人が集まりました。</td>
<td>The event was organized by the private Foundation for Global Peace and Environment.</td>
</tr>
<tr>
<td>参加者たちは裸足になって水田に足を踏み入れ地元に伝わる田植え歌に合わせて慣れない手つきで苗を植えていました。</td>
<td>The rice seedlings are grown from grain harvested by the Emperor at the Imperial Palace in Tokyo last autumn.</td>
</tr>
<tr>
<td>きょうの輪島市は吉祥が広がったものの月曜の天気となり、出席された高円宮さまも海からの風に吹かれながら田植えに加わっていました。</td>
<td>Barefoot participants waded into the paddies to plant the seedlings by hand while singing a local folk song about the practice of rice planting.</td>
</tr>
</tbody>
</table>
Phrase Alignment

- Phrase-alignment in 1-to-1 aligned-sentence-pair
- We use our method proposed in [Aramaki et al. 2001, MT-Summit VIII]
  - 1: Analysis of phrasal dependency structures
  - 2: Estimation of basic-phrasal corresp. with dictionaries
  - 3: Expansion of phrasal corresp. with surrounding information

The rice seedlings are grown from grain harvested by the Emperor at the Imperial Palace in Tokyo last Autumn.

苗は (rice seedling)
去年の (of last year)
秋 (autumn)
天皇陛下が (the Emperor)
皇居で (at the Imperial Palace)
収穫された (harvested)
稲穂から (from the chaff of rice)
育てられたものです (grown)
Translation Example (TE) and Translation Memory (TM)

- TE := sentence pairs
  - Structurally analyzed
  - Aligned at phrase level

- TM := a collection of TEs

Translation Memory =

The rice seedlings are grown from grain harvested by the Emperor at the Imperial Palace in Tokyo last Autumn.

苗は(稲の苗)去年の(昨年の)秋(秋)は、(天皇陛下が)(the Emperor)皇居で(at the Imperial Palace)収穫された(稲の苗)、(from the chaff of the rice)育てられたものです。
Evaluation of Alignment Accuracy

- **Sentence alignment**
  - Evaluation-Data: 96 article pairs
  - Result:
    - All(m-to-n) alignment = 60% (=226/377)
    - 1-to-1 alignment = 77% (=111/145)

- **Phrase alignment**
  - Evaluation-Data: 145 1-to-1 sentence pairs
  - Result:
    - Phrase alignment precision= 50%
    - not high enough to use for TM
WCR (Word Corresponding Ratio)

- Discard sentence-pairs with little word correspondence

$$WCR = \frac{\# \text{ Content words corresponded in dictionaries}}{\# \text{ Content words}}$$

Example:

- Japanese: 秋, 天皇陛下が, 去年の収穫された稲穂から, 育てられたものです.
- English: The rice seedlings are grown from grain harvested by the Emperor at the Imperial Palace in Tokyo last Autumn.

$$WCR = \frac{6 + 8}{9 + 11} = 0.7$$
WCR & Phrase Alignment Precision

30% 30,000 TEs

70,000 TEs
Outline

1: NHK News Corpus

2: How to build a Translation Memory (TM)

3: How to select Translation Examples from TM

4: Experiments

5: Conclusion
Translation Algorithm

Input Sentence  TE  Output Sentence
Question

- How do we select the most plausible example?

- **Equality** between ı and Ş
- Surrounding **Similarity** of ı and Ş
- Alignment **confidence** between Ş and T
The United States has issued a request that restricts exports.
1: Equality between I & S

- Equality := # of equal phrases in I & S

Equality between I & S and S:

- The United States has issued a request that restricts exports.
- This is reflected in the equality score.

Diagram:

- Congress
- Import
- Restrict
- Recommend
- America
- Export
- Restrict
- Recommend

TE Score = EQ
2: Surrounding Similarity

The United States has issued a request that restricts exports to America. The Thesaurus match score is 0.3 to 0.8, and the POS match score is 0.3.

TE Score = EQ + SIM
Confidence of Alignment

- is estimated by using dictionaries
- is estimated by surrounding information

\[
\text{TE Score} = \left\{ \square \text{EQ} + \square \text{SIM} \right\} \square \square \text{CONF}
\]
Global Confidence (WCR)

The United States has issued a request that restricts exports.

TE Score = \{ \text{EQ} + \text{SIM} \} \text{ CONF WCR}
Example
Outline

1: NHK News Corpus

2: How to build a Translation Memory (TM)

3: How to select Translation Examples from TM

4: Experiments

5: Conclusion
Experiments (word selection task)

- **Evaluation-Data**
  - 50 Japanese sentences in NHK-corpus (not used for TM)

- **Gold-standard-Data**
  - 50 English sentences (which are pairs of evaluation data)

- **Evaluation**
  - A human judge phrase by phrase referring gold-standard-data

<table>
<thead>
<tr>
<th>Method</th>
<th>EQ</th>
<th>SIM</th>
<th>CONF</th>
<th>Good</th>
<th>Bad</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>EQ</td>
<td>SIM</td>
<td>CONF</td>
<td>268</td>
<td>47</td>
<td>85.0%</td>
</tr>
<tr>
<td>Method A</td>
<td>EQ</td>
<td></td>
<td>CONF</td>
<td>254</td>
<td>61</td>
<td>80.6%</td>
</tr>
<tr>
<td>Method B</td>
<td>EQ</td>
<td>SIM</td>
<td></td>
<td>234</td>
<td>81</td>
<td>74.2%</td>
</tr>
<tr>
<td>Baseline</td>
<td>Dic.</td>
<td>Freq.</td>
<td></td>
<td>232</td>
<td>83</td>
<td>73.6%</td>
</tr>
</tbody>
</table>

Baseline method uses only dictionaries, and selects the most frequent word or phrase.
Example

望んでいます
(joined) have been welcomed

Input Sentence

キム デジュン大統領は、
(President Kim Dae – jung) The Japanese Emperor and Empress
現在
(now) 昨夜
歓迎晩餐会に
(at reception dinner) (at the ceremony)
臨んでいます．
(joined) have been welcomed

TE

望んでいます
(joined) have been welcomed

The Japanese Emperor Empress
have been welcomed
at a ceremony

- TE has much similarity to the Input sentence.
- TE has low alignment confidence.
Outline

1: NHK News Corpus

2: How to build a Translation Memory (TM)

3: How to select Translation Examples from TM

4: Experiments

5: Conclusion
Conclusion

- EBMT system using a content-aligned corpus
- Proposed methods:
  - TM Construction
    - Discard sentence-pairs with too little word correspondence
    - Rigorous phrase alignment
  - TE selection
    - Source language similarity
    - Translation confidence
  - The accuracy of the word selection was 85%
- Future Work:
  - Completion of remaining components
  - Evaluation of full Translations