Bilingual parallel corpora are an extremely important resource as they are typically used in data-driven machine translation. There already exist many freely available corpora for European languages, but almost none between Chinese and Japanese. The constitution of large bilingual corpora is a problem for less documented language pairs. We construct a quasi-parallel corpus automatically by using analogical associations based on certain number of parallel corpus and a small number of monolingual data. Furthermore, in SMT experiments, by adding this kind of Chinese–Japanese data into the baseline training corpus, on the same test set, the evaluation scores of the translation results we obtained were significantly or slightly improved over the baseline systems.

### Building analogical clusters according to proportional analogies

- **Proportional analogy** establishes a general relationship between four objects \(A, B, C\) and \(D\): \(A : B = C : D\). An efficient algorithm for the resolution of analogical equations has been proposed in (Lepage, 1998).

\[
\begin{align*}
|A_B| &= |B| - |C| - |D|, \forall n \\
d(A, B) &= d(C, D) \\
d(A, C) &= d(B, D)
\end{align*}
\]

- **Sentential analogy:**

  - Early in the process, we set up sentential analogies as a sequence of lines, where each line contains one sentence pair and where any two pairs of sentences form a sentential analogy.

  - Early in the process, we set up sentential analogies as a sequence of lines, where each line contains one sentence pair and where any two pairs of sentences form a sentential analogy.

- **Analogue cluster:**

  - We can construct sentential analogies as a sequence of lines, where each line contains one sentence pair and where any two pairs of sentences form a sentential analogy.

  - We can construct sentential analogies as a sequence of lines, where each line contains one sentence pair and where any two pairs of sentences form a sentential analogy.

- **We produced all possible analogical clusters from Chinese and Japanese unrelated unaligned monolingual data collected from the Web.**

<table>
  <tr><th>Chinese</th><th>Japanese</th></tr>
  <tr><td># of different sentences</td><td>70,000</td><td>70,000</td></tr>
  <tr><td># of clusters</td><td>23,192</td><td>21,376</td></tr>
</table>

Such clusters can be considered as rewriting models that can generate new sentences.

### Extracting corresponding clusters by computing similarity according to a classical Dice formula:

\[
\text{Sim}_{\text{ Dice}} = \frac{|S_a \cap S_b|}{|S_a \cup S_b|} = \frac{\text{sim}_{text}(A, B) + \text{sim}_{vocab}(A, B)}{|S_a \cup S_b|}
\]

where \(S_a\) and \(S_b\) denote the minimal sets of changes across the clusters (both on the left or right) in both languages (after translation and conversion).

### Generation of new sentences using analogical associations

- **Generation of new sentences**

  We use analogy as an operation by which, given two related forms (rewriting model) and only one form, the fourth missing form is coined. Applied on sentences, this principle can be illustrated as follows:

  \[
  \text{早後に対応して下さい。} \quad \rightarrow \quad \text{早後に対応して下さい。}
  \]

- **Experiments on new sentence generation and filtering by N-sequences**

  We eliminate any sentence that contains an N-sequence of a given length unseen in our data. For valid sentences, we remember their corresponding seed sentences and the cluster identifiers they were generated from.

### Deducing and acquiring quasi-parallel sentences

We deduce translation relations based on the initial parallel corpus and corresponding clusters between Chinese and Japanese.

### Experimental results (using the different segmentation tools and Moses version):

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Chinese</th>
<th>Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>6.9533</td>
<td>7.3208</td>
</tr>
<tr>
<td>NIST</td>
<td>0.5633</td>
<td>0.5667</td>
</tr>
<tr>
<td>WER</td>
<td>0.5427</td>
<td>0.5667</td>
</tr>
<tr>
<td>TER</td>
<td>0.5427</td>
<td>0.5667</td>
</tr>
</tbody>
</table>

| BLEU     | 7.0422  | 7.3123   |
| NIST     | 0.5427  | 0.5852   |
| WER      | 0.5427  | 0.5852   |
| TER      | 0.5427  | 0.5852   |

Such results are using: segmentation tools: kytea, moses 1.0 to 1.1.