Sentence pattern generation for low-resource language pairs

VARGA István, YOKOYAMA Shoichi, HASHIMOTO Chikara

Yamagata University

Introduction

Highly accurate, fast and economically affordable machine translation systems remain an elusive goal in MT. Although due to the various translation methods and refinements there are a number of success stories, the uncovered territory is still dominant. In the case of low-resource languages or less-common language pairs the question is far more complex: the importance of choosing the most appropriate technique is eclipsed by the resource limitations, whether that’s manifested by the lack of personnel or machine translation tools for the languages in question. To overcome these problems, economical and efficient tools and methods are needed.

Research purpose

The goal of our research is an economically viable Hungarian-Japanese MT system for gisting purposes. We propose MT techniques that can be reproduced with virtually every languages pair, assuming that there are some monolingual or bilingual tools, however limited. In this current study we are designing a method to automatically generate sentence patterns and grammatical rules, concentrating on low resource languages. Our method assumes the existence of parsers for both languages, as well as the availability of a small bilingual corpus.

Related works

There are already numerous methods for sentence pattern generation. The are a limited number of success stories (Alintas & Güvenir, 2003; Gedeck, 2005), but these methods only work with closely related languages, since they do not need any deep grammatical analysis. Other automated methods make use of large bilingual corpora and a bilingual dictionary, looking for some sort of structural similarity between the counterpart sentences (Watanabe et al., 2000; Kaji et al., 1992). However, most of these methods still clude the desired accuracy. We believe that the main reason for this is that these methods work with separate sentence pairs, trying to extract sentence patterns for each pair. Specially with distant languages, whose different grammatical structure offers no help, these methods produce many erroneous, useless or even contradictory results.

Step 1: corpus acquisition

There is no known digital bilingual corpus between Japanese and Hungarian. We built the following corpora:

<table>
<thead>
<tr>
<th>text type</th>
<th>size (approx. sentence pairs)</th>
<th>acquisition method</th>
<th>comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>translated</td>
<td>6500</td>
<td>free download</td>
<td>1-to-1 rare, structurally inconsistent</td>
</tr>
<tr>
<td>literature (3rd language)</td>
<td>15000</td>
<td>scan</td>
<td>structurally inconsistent</td>
</tr>
<tr>
<td>translated</td>
<td>30000</td>
<td>free download</td>
<td>noisy translations, probably MT</td>
</tr>
<tr>
<td>literature (direct)</td>
<td>3000</td>
<td>manual typing</td>
<td>many short sentences, but grammatically rich</td>
</tr>
<tr>
<td>help files</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>language books</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Step 2: grammatical profile generation

We exemplify our method with a 4-sentence bilingual corpus:


We generate the “grammatical profile” of each language, accumulating all parse trees into one. We are looking for frequent patterns in each language.

Step 3: pattern generation

Method characteristics:

- analyze frequent patterns, instead of each sentence pair
- process from both sides (source and target)
- generate from subtrees, instead of full parsed trees
- bottom-to-top processing: start from the lowest unprocessed level
- generate the most general patterns; subcategories as needed

Example:

"H": s1.6. V (V + "は") - V (freq=1) (rule6) V - AV (from dict)
"J": s2.7. V (V + "は") - V (freq=1) (rule6) V - AV (from dict)

"H": s1.6. V (V + "は") - V (freq=1) (rule6) V - AV (from dict)
"J": s2.7. V (V + "は") - V (freq=1) (rule6) V - AV (from dict)

"H": s1.6. V (V + "は") - V (freq=1) (rule6) V - AV (from dict)
"J": s2.7. V (V + "は") - V (freq=1) (rule6) V - AV (from dict)

References: