Development of an Asymptotic Word Correspondence System between Classical Japanese Poems and their Modern Translations

Hilofumi Yamamoto†‡
* University of California, San Diego †Tokyo Institute of Technology

Introduction

• This project will develop an automatic word concordance system for parallel texts comprising of Classical Japanese poems and their associated modern translations.
• By using these parallel texts, we will clarify the details of language change within Japanese in an objective procedural manner that is not influenced by human observations.
• Our aim is to develop a thesaurus of classical Japanese poetic vocabulary using the system.

Problem

What is Waka?

1. Orthography Problem
2. Unit size Problem
3. Attribution Problem
4. Polysemy/PUN Problem

Methods

Material: Kokinshū a.k.a. Kininwakashū is:
the first anthology compiled by the order of Emperor Daigo (ca. 905), which contains about 1,100 poems. And 10 sets of their Contemporary Japanese Translations (CT)

Kokinshū (905)
Kokinwakashū (906)
Sanka (1609–15)
Sanka (1415–16)
Mimosa (1897)
Mimosa (1890)
Ozawa (1971)
Kubota (1960)
Katagiri (1998)

Figure 1: Dates of publication of annotations of the Kokinshū:
• indicates that it has CT; ● indicates that it does not include CT; ▲ indicates that it is used in this project.

Result

Mutual Co-occurrence Rate: Murai (2010)

\[ \text{mcr}(o, t) = p(o \mid t) \cdot p(t \mid o) \]

where, \( o \) indicates a token in original texts; \( t \), a token in translation texts; \( mcr(o, t) \), the mutual co-occurrence rate; \( p(o \mid t) \), the rate when a token \( o \) and \( t \) occur at the same time in corresponding texts which are original texts and translation texts.

→ when mcr is large enough, it will be estimated that token \( o \) and \( t \) are contextually equivalent.

Good or poor estimated pairs

Table 1: Good estimated pairs and poor estimated pairs; the values of good pairs are the first ten items (over 1.3); and the values of poor pair items are the last ten items (lower 0.01).

<table>
<thead>
<tr>
<th>no.</th>
<th>good pairs</th>
<th>poor pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>鳴く</td>
<td>鳴く</td>
</tr>
<tr>
<td>2</td>
<td>風</td>
<td>風</td>
</tr>
<tr>
<td>3</td>
<td>世の中</td>
<td>世の中</td>
</tr>
<tr>
<td>4</td>
<td>人</td>
<td>人</td>
</tr>
<tr>
<td>5</td>
<td>春</td>
<td>春</td>
</tr>
<tr>
<td>6</td>
<td>秋</td>
<td>秋</td>
</tr>
<tr>
<td>7</td>
<td>時鳥</td>
<td>時鳥</td>
</tr>
<tr>
<td>8</td>
<td>時鳥</td>
<td>時鳥</td>
</tr>
<tr>
<td>9</td>
<td>散る</td>
<td>散る</td>
</tr>
<tr>
<td>10</td>
<td>見る</td>
<td>見る</td>
</tr>
</tbody>
</table>

Conclusion

1. This project has already begun: the parallel corpus of the Kokinshū has been constructed.
2. We are now working on the development of computer software and the optimization of the calculation methods.

Reference