Symmetric Correspondence Topic Models for Multilingual Text Analysis

Kosuke Fukumatsu (Kobe University)
Koji Eguchi (Kobe University)
Eric P. Xing (Carnegie Mellon)
Multi-type topic models

- Topic modeling approach has been applied to multi-type data, such as:
  - text-annotated image data
  - multilingual aligned text data

- Some prior models fail to capture dependencies between the data types.

- CorrLDA (Correspondence LDA) [Blei et al., 2003] does well with the dependencies; however, it requires an annoying constraint, which requires a *pivot* — which data type is generated first — to be specified in advance.
Symmetric Correspondence LDA

- **SymCorrLDA** can adjust the proportion of the pivot for each data type, using a switch variable.
We experimented with the multilingual aligned text data gathered from *Wikipedia*.

- CorrLDA works significantly more effectively than the other prior models.
- SymCorrLDA is more effective than CorrLDA.

**Results**

Mean reciprocal rank in finding counter parts.

<table>
<thead>
<tr>
<th></th>
<th>Jap to Eng</th>
<th>Eng to Jap</th>
</tr>
</thead>
<tbody>
<tr>
<td>CorrLDA</td>
<td>0.299</td>
<td>0.316</td>
</tr>
<tr>
<td>SymCorrLDA</td>
<td><strong>0.326</strong></td>
<td><strong>0.335</strong></td>
</tr>
</tbody>
</table>

9% improved. 6% improved. (Both improvements are statistically significant.)