HMM Taggers

- Corpus statistics provided by Chih-Hao Chang
Training Corpus

• Number of Sentences : 697,712

• With punctuation markers
  – Number of Words (Tags) : 6,027,509
  – Number of Distinct Words : 126,782
  – Number of Distinct Tags : 57

• Without punctuation markers
  – Number of Words (Tags) : 4,485,838
  – Number of Distinct Words : 126,529
  – Number of Distinct Tags : 47
Testing Corpus

- Number of Sentences : 15,038

- With punctuation markers
  - Number of Words (Tags) : 147,538
  - Number of Distinct Words : 13,196
  - Number of Distinct Tags : 56

- Without punctuation markers
  - Number of Words (Tags) : 113,594
  - Number of Distinct Words : 13,146
  - Number of Distinct Tags : 45
An Example Result

- HMM Tagger using Bi-gram Information
- Preliminary Result on the Testing Corpus

<table>
<thead>
<tr>
<th></th>
<th>Error Tokens</th>
<th>Total Tokens</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without Punctuation Markers</td>
<td>6,074</td>
<td>1,135,94</td>
<td>94.65</td>
</tr>
<tr>
<td>With Punctuation Markers</td>
<td>5,758</td>
<td>147,538</td>
<td>96.09</td>
</tr>
</tbody>
</table>