Overview

- **Objective**: Create a pronunciation estimator that is robust and adaptable to new domains.
- **Focus on Japanese/Chinese, where we must estimate word boundaries as well**.
- **Approach**: Pointwise prediction, which tags all word boundaries and pronunciations independently.
- **Pointwise prediction**:
  - **Robust**: relies on dictionaries less than previous methods.
  - **Adaptable**: it can be learned from single annotated words, not full sentences.
- **Evaluation**: on Japanese pronunciation estimation shows improvement over traditional joint n-gram.

Pronunciation Estimation Methods

- **Joint**: Predict word boundaries+tags simultaneously.
- **Joint n-gram language models are representative**.
- **2-Step**: First predict word boundaries, then pronunciations.
- **Can use Logistic Regression, SVM, CRF**.
- **LR and SVM are pointwise, CRF not**.

Features for Pointwise Prediction

- **Specify features using character n-grams, character type n-grams, length-annotated dictionary presence**.
- **Boundary Point**
  - WS
  - Pron
  - One pronunciation: Use it
  - Unknown word: Character-based noisy channel model
  - Multiple candidates: Use classifier

Key point: None of the features require word boundaries or surrounding tags.

Annotation Methods

- **Pronunciation estimation underperforms on out-of-domain text → would like to adapt efficiently**.
- **Necessary to create data in the target domain**.
- **Resources for training joint predictors**:
  - Word/Pron Corpus
  - Word/Pron Dictionary
  - Partially Annotated Corpus
  - Word Dictionary

Experiments

- **Experiments performed on two domains**.
- **General domain**: Balanced corpus of contemporary written Japanese (BCCWJ).
- **Target domain**: Business newspaper text.
- **Tested two systems**.
  - Joint trigram model.
  - Pointwise: The 2 step method using linear SVMs.
- **Evaluation**: Mora-based precision and recall.

<table>
<thead>
<tr>
<th></th>
<th>Joint Trigram</th>
<th>Pointwise</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Test</td>
</tr>
<tr>
<td>General</td>
<td>99.07%</td>
<td>99.12%</td>
</tr>
<tr>
<td>Target</td>
<td>97.83%</td>
<td>97.23%</td>
</tr>
</tbody>
</table>

- **Pointwise approach uniformly better**.
- **Domain adaptation using partial annotation**.
- **Partially annotated corpus (PAC):**
  - Annotated 1,366 words in target-domain training corpus with boundaries and pronunciations.
  - Chosen according to classifier accuracy.
- **Dictionary of word sequences (DWS):**
  - Chose 1,060 words from a compound word dictionary that were most frequent in training corp.
  - Annotated with words and pronunciations.

<table>
<thead>
<tr>
<th></th>
<th>Joint Trigram</th>
<th>Pointwise</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Test</td>
</tr>
<tr>
<td>General</td>
<td>97.83%</td>
<td>97.23%</td>
</tr>
<tr>
<td>Target</td>
<td>98.02%</td>
<td>97.51%</td>
</tr>
<tr>
<td>Change</td>
<td>+0.29%</td>
<td>+0.28%</td>
</tr>
</tbody>
</table>

Available Open Source!
http://www.phontron.com/kytea/