**Overview**

- **Objective:** Create a Japanese morphological analyzer (word segmentation + POS tagging) that is robust and adaptable to new domains
- **Approach:** Use pointwise prediction, which estimates all tags independently of other tags
- **Pointwise prediction:**
  - **Robust:** does not rely on dictionaries as much as previous methods
  - **Adaptable:** it can be learned from single annotated words, not full sentences
  - **Works with active learning:** Single words to annotate can be chosen effectively
- **Evaluation** on Japanese morphological analysis shows improvement over traditional methods

**Features for Pointwise MA**

- **Specify features using character n-grams, character type n-grams, length-annotated dictionary presence**

**Annotation Methods**

- Morphological analysis underperforms on out-of-domain text → we would like to adapt
- We have an in-domain unannotated text, and some annotator time
- **Goal is to maximize the effect** for annotator time
- **Use active learning** to choose data to annotate

**Morphological Analysis Methods**

- **Joint:** Predict word boundaries+tags simultaneously
- **Use HMMs, CRFs, or language models**

- **2-Step:** First predict word boundaries, then POSs
  - Can use Logistic Regression, SVM, CRF
  - LR and SVM are pointwise, CRF not

**Experiments**

- **Experiments performed on the Balanced Corpus of Contemporary Written Japanese (BCCWJ)**
- **General domain:** News, white papers, books
- **Target domain:** Web text

**Experiments**

- Tested three systems
  - **Joint:** Kudo et al.’s CRF-based method, as implemented by the MeCab toolkit
  - **2-CRF:** The 2 step method using CRFs as a solver
  - **2-LR:** 2 step pointwise method using LR

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>97.31% 98.08% 98.03%</td>
</tr>
<tr>
<td>General</td>
<td>94.57% 95.39% 95.13%</td>
</tr>
<tr>
<td>Gen+Tar</td>
<td>96.45% 96.91% 96.82%</td>
</tr>
</tbody>
</table>

- **2-LR slightly worse than 2-CRF, better than Joint**
- Tested 2 annotation strategies for domain adaptation
- **Sentence:** annotate the sentence with lowest overall posterior probability
- **Partial:** annotate the word with lowest prob. margin

- **Available Open Source!**
  - http://www.phontron.com/kytea/
  - Chinese models, Japanese pronunciation estimation also available