Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis

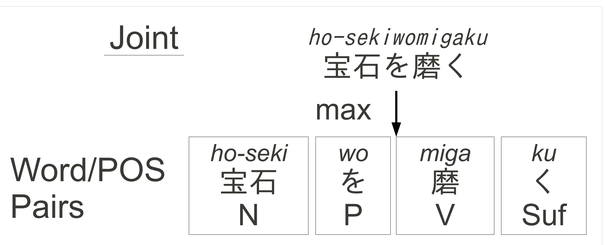
Graham Neubig, Yosuke Nakata, Shinsuke Mori
Graduate School of Informatics, Kyoto University, Japan

Overview

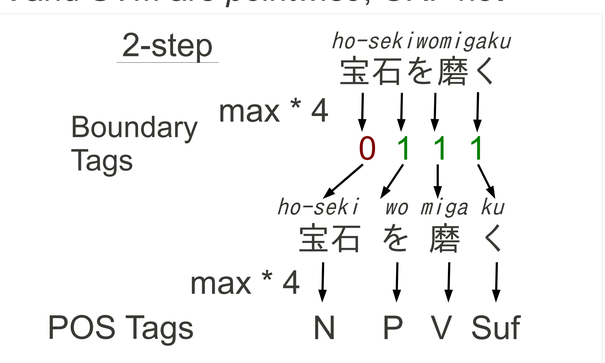
- **Objective:** Create a Japanese morphological analyzer (word segmentation + POS tagging) that is robust and adaptable to new domains
- **Approach:** Use pointwise prediction, which estimates all tags independently of other tags
- **Pointwise prediction:**
  - *Robust:* does not rely on dictionaries 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

Morphological Analysis Methods

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

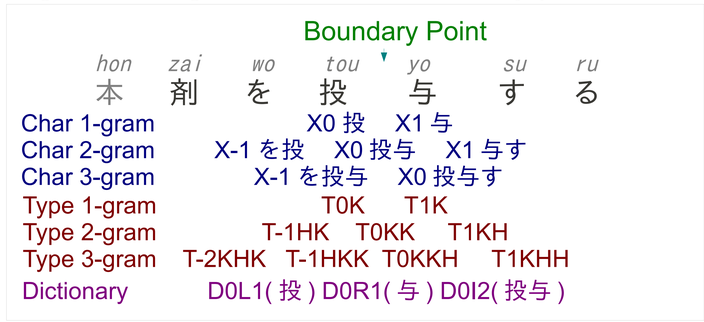


- **2-Step:** First predict word boundaries, then POSs
  - Can use Logistic Regression, SVM, CRF
  - LR and SVM are *pointwise*, CRF not



Features for Pointwise MA

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



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

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

Reference
本剤/N を/P 投与/N する/V ため/N
Automatic Result

本/Pre	剤/N	を/P	投/N	与/N	する/V	ため/N
0.8	0.91	1.0	0.98	0.94	0.998	0.997

- **Full annotation:** Choose sentences with low prob.
  - Can train any model on this annotated data

本剤/N を/P 投与/N する/V ため/N
Annotated (5)

- **Partial annotation:** Choose words with low prob.
  - Only pointwise prediction can be used

本剤/N を 投与/N するため
Annotated (2) Unannotated

Experiments

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

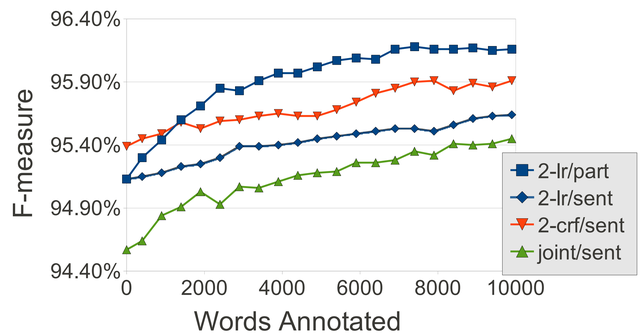
	Train	Test
General	782k	87.5k
Target	153k	17.3k

- 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

Train	Test	Joint	2-CRF	2-LR
General	General	97.31%	98.08%	98.03%
General	Target	94.57%	95.39%	95.13%
Gen+Tar	Target	96.45%	96.91%	96.82%

- 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



- Partial annotation much more effective

Available Open Source!

http://www.phontron.com/kytea/

Chinese models, Japanese pronunciation estimation also available