Overview

- **Objective**: Minimize the amount of effort required for domain adaptation
- **Approach**: A word-based partial annotation strategy, and a machine-learning strategy that can utilize partially annotated data
- **Details**:
  - Use a point-wise classifier to allow for learning from partially annotated data
  - Introduce a strategy to pick annotation segments based on character bi-gram diversity
- **Evaluation** on word segmentation and pronunciation estimation for Japanese shows improvement over full annotation

Japanese Pronunciation Estimation

- Consists of two elements
  - **Word segmentation (WS)** that divides unsegmented characters into words
  - **Pronunciation Estimation (PE)** that finds the appropriate pronunciation for each word

![Diagram showing Word Segmentation and Pronunciation Estimation](image)

- Previously proposed methods use sequence-based estimation (e.g. n-gram models)

Language Resources

**General Domain**

- **Balanced Corpus of Contemporary Written Japanese (BCCWJ)**: 898k words fully annotated with pronunciations and word boundaries
- **UniDic**: 212k word dictionary annotated with 1.05 pronunciations/word
- **Number Dictionary**: Dictionary of 2 and 4-digit numbers with pronunciations for use in years

**Target Domain**

- Creating language resources in the target domain will increase accuracy
- Difficult and time consuming!

Partial Annotation

- Most target domain sentences only contain a few points not covered by the general domain resources
- **Full annotation wastes time on well covered points!**
- **Solution**: Only annotate points that are not well covered in the general domain

![Diagram showing Word Segmentation and Pronunciation Estimation](image)

- Character bigrams that exist in the target corpus but not the general corpus were selected (in order of frequency)

Point-Wise Estimation

- **Traditional sequence-based (n-gram) methods cannot learn from partial annotation!**
- **Solution**: Use point-wise estimation, which estimates each word boundary or pronunciation independently of the others
- Estimation is performed using linear SVMs or logistic regression
- Features used:
  - Character n-gram, character type n-gram, dictionary words

Available open-source: [http://www.phontron.com/kytea](http://www.phontron.com/kytea)

Experimental Results

- **Target domain**: Nikkei business newspaper
- **Training**: 263k words, **Test**: 29k words
- **Estimation strategy**: Tri-gram vs. Point-wise
- **Annotation strategies**: Full vs. Partial annotation
- **Results**: Point-wise partial approach most effective

![Graph showing Word Segmentation and Pronunciation Estimation](image)