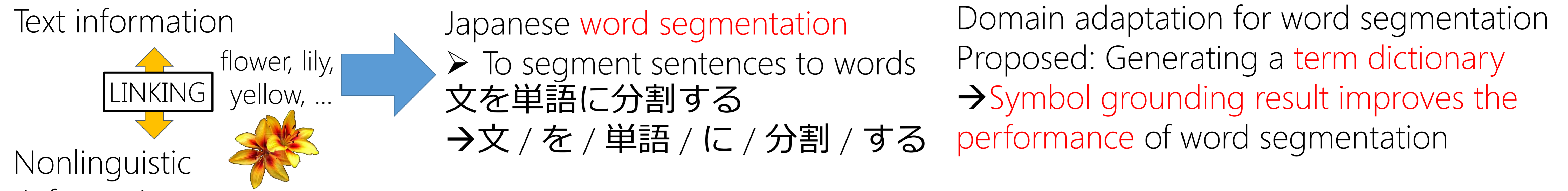


# Can Symbol Grounding Improve Low-Level NLP?

## Word Segmentation as a Case Study

Hiroataka Kameko<sup>†</sup>, Shinsuke Mori<sup>‡</sup> and Yoshimasa Tsuruoka<sup>†</sup>  
<sup>†</sup>The University of Tokyo, {kameko, tsuruoka}@logos.t.u-tokyo.ac.jp  
<sup>‡</sup>Kyoto University, forest@i.kyoto-u.ac.jp

### Symbol Grounding for Text Associated with Multi-Modal Information



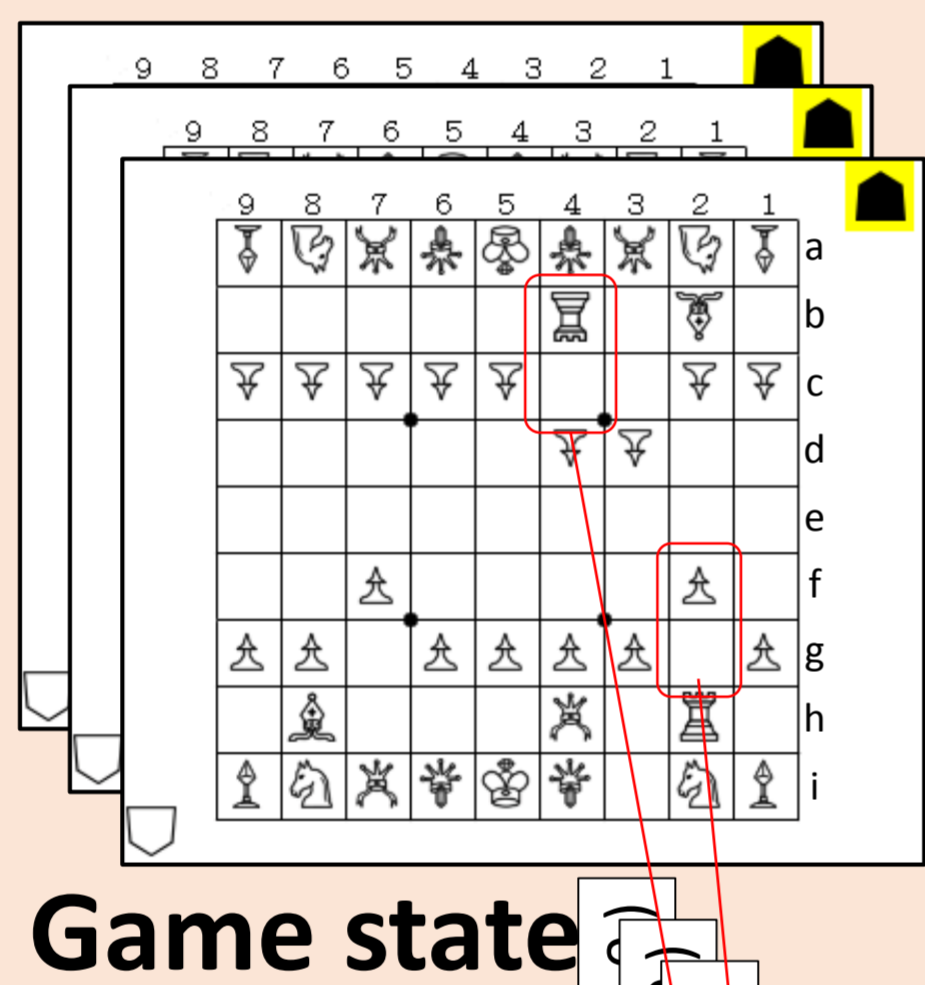
### Method: Automatic Term-Dictionary Generation

Commentary for Shogi (Japanese chess)  
 □ Written in Japanese  
 □ Many terms associated to game states  
 Corpus: a set of pairs of a Shogi state and a comment

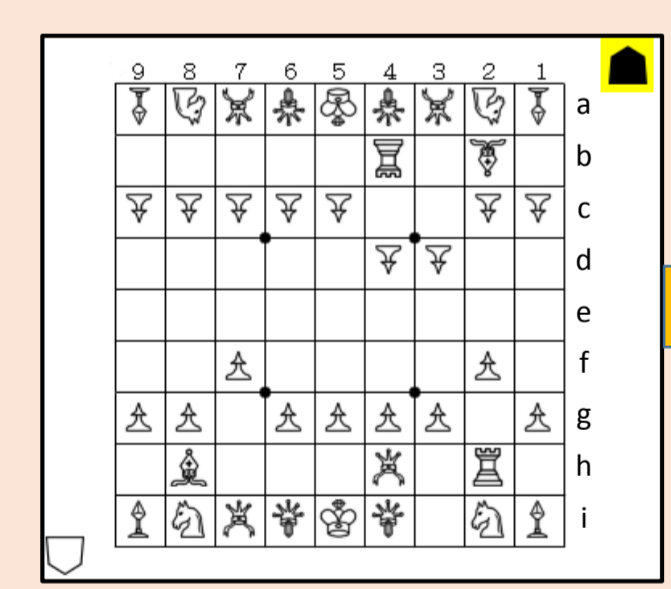
#### 2. Symbol Grounding

Training multi-layer perceptron to choose terms for the state with collect fragments

Terms for the state:  
 Terms and features of the states have strong correlations  
 Collect fragments:  
 Candidates with wrong fragments appear randomly



- Input: features of Shogi states
- a) positions of pieces
  - b) pieces captured
  - c) combinations of a) & b)
  - d) other heuristics

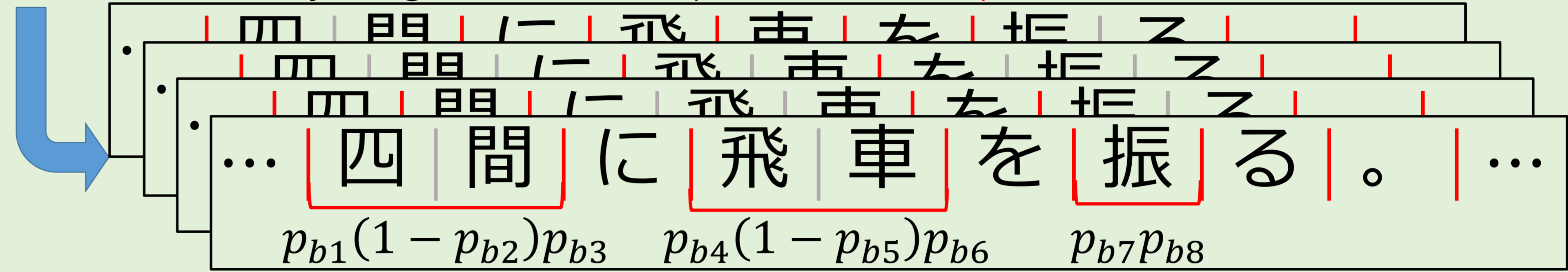


四間|に|飛|車|を|振|る|  
 0.8 0.1 0.2 0.1 0.05 0.2  
 四間|に|飛|車|を|振|る|  
 0.6 0.002 0.7 0.05 0.4 0.3  
 四間|に|飛|車|を|振|る|  
 0.001 0.3 0.002 0.4 0.3

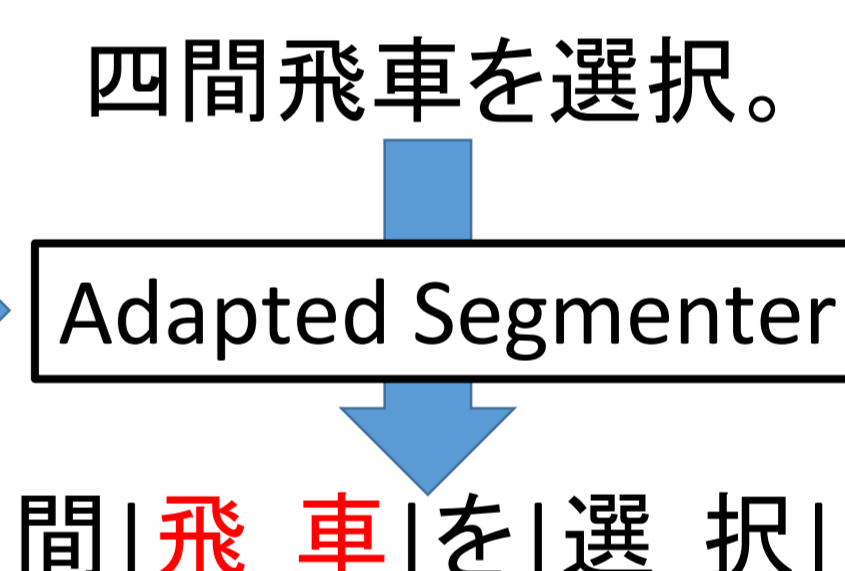
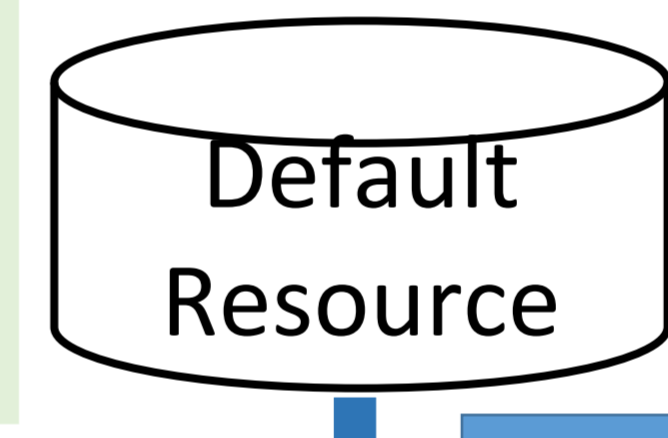
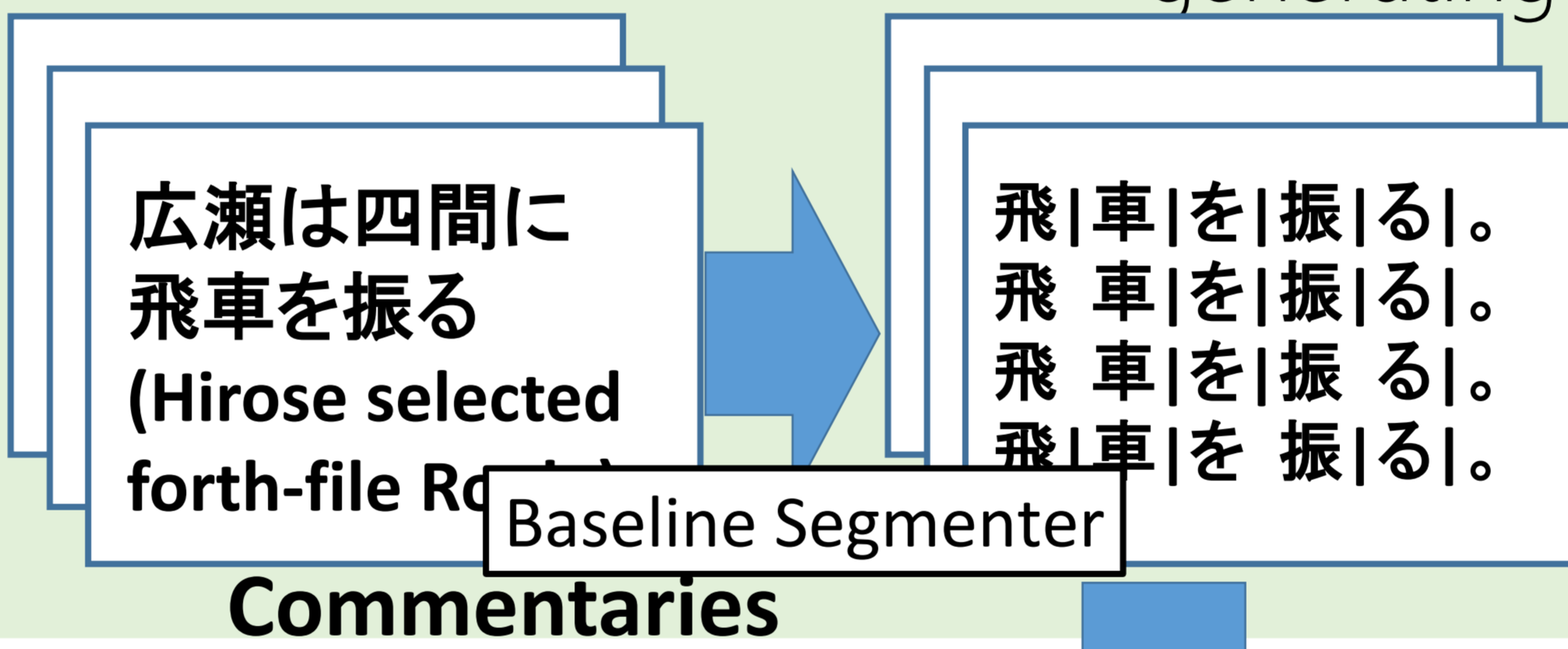
1. pseudo-Stochastically Segmented Corpora [Mori and Takuma, 2004]  
 $p_{bi}$ : probability of boundary (by baseline segmenter)

$p_{b1} p_{b2} p_{b3} p_{b4} p_{b5} p_{b6} p_{b7} p_{b8} p_{b9} p_{b10}$   
 ... | 四 | 間 | に | 飛 | 車 | を | 振 | る | 。 | ...

Stochastically Segmented Corpora include probable candidate words



generating deterministically segmented corpora (in this experiment, 4 times)



Output: words which will appear in the comments for the states

飛 (fly)  
 車 (chariot)  
 飛車 (Rook)  
 を

higher scores:  
 terms for the state & collect fragments

#### 3. Dictionary Generation

Selecting word candidates by summation of the scores

Top  $N$  word candidates  
 Tuning  $N$  by measuring the accuracies on the development set  
 In this experiment,  $N = 127$

### Evaluation

#### Corpus specifications

|             |              | # of sent. | # of words | # of char. |
|-------------|--------------|------------|------------|------------|
| Training    | BCCWJ        | 56,753     | 1,324,951  | 1,911,660  |
|             | Newspaper    | 8,164      | 240,097    | 361,843    |
|             | Conversation | 11,700     | 147,809    | 197,941    |
| Development | Shogi-dev.   | 170        | 2,501      | 3,340      |
| Test        | BCCWJ-test   | 6,025      | 148,929    | 212,261    |
|             | Shogi-test   | 3,299      | 24,966     | 32,481     |

#### Accuracy on Shogi commentaries

|             | Recall | Precision | F-Measure |
|-------------|--------|-----------|-----------|
| Baseline    | 90.12  | 91.43     | 90.77     |
| + Sym. Gro. | 90.60  | 91.66     | 91.13     |

Target domain: Our framework successfully acquired new words

#### Accuracy on BCCWJ

|             | Recall | Precision | F-Measure |
|-------------|--------|-----------|-----------|
| Baseline    | 98.99  | 99.06     | 99.03     |
| + Sym. Gro. | 99.03  | 99.01     | 99.02     |

General domain: Our framework did not cause a severe performance degradation

### Conclusion

- Symbol grounding can improve word segmentation
- The method requires a small amount of annotated data
  - only to tune the hyperparameter
- The framework is general: It is worth testing on other tasks

### Future Work

- To apply our approach to other tasks
- To deal with other types of non-textual information
  - e.g.) images, economic indices