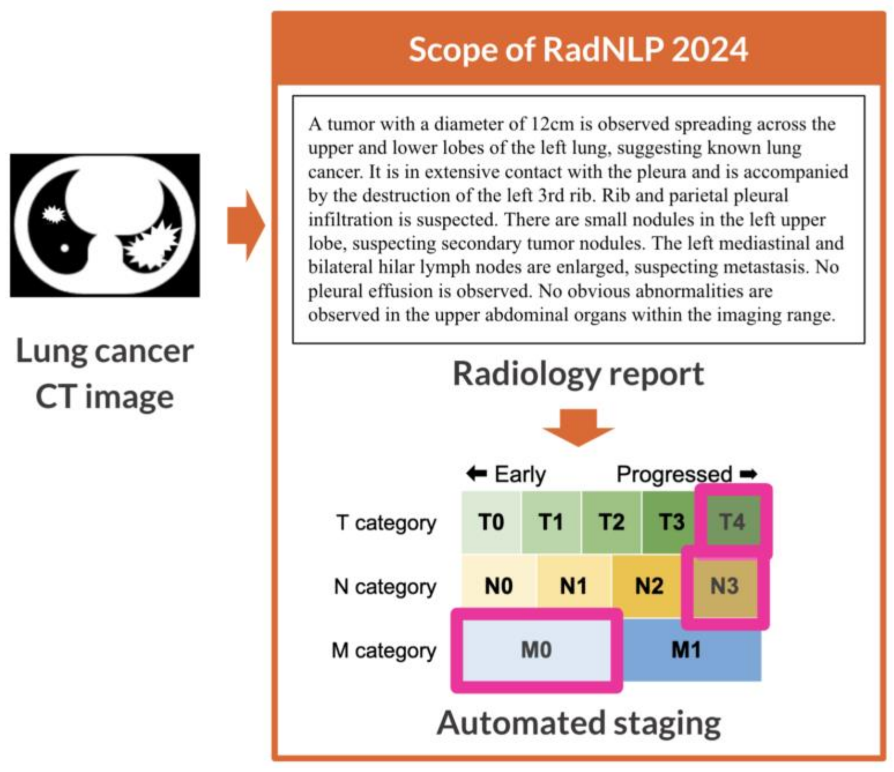
Domain Adaptation with Medical Vocabulary-Aware Tokenizer for Radiology Report Analysis in RadNLP at KAIYO03

Daiki Shirafuji[1] Takafumi Niwa[2]

[1]Mitsubishi Electric Corporation [2]National Defense Medical College

Explanation: Sub Task on RadNLP



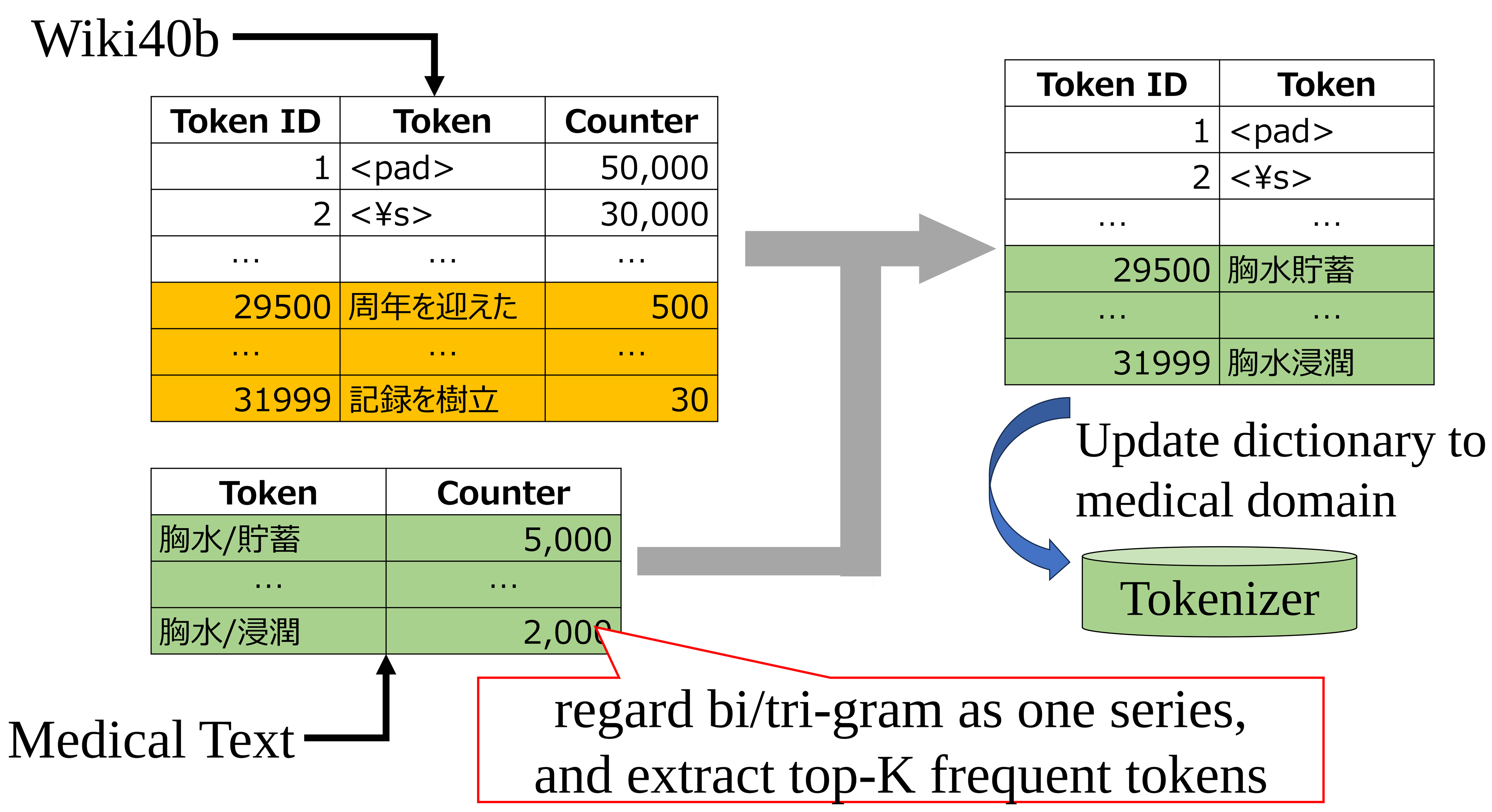
RadNLP sub task: the eight-label sentence binary classification. (Sentence-level segmentation)

Measure: Span describing mainly the existence and diameter of the primary lesion.
Extension: Extent of the primary lesion's spread beyond the lung parenchyma.
Atelectasis: Span pointing out atelectasis or obstructive pneumonia.
Satellite: Span pointing out intrapulmonary metastasis or lymphangiomatosis carcinomatosa.
Lymphadenopathy: Span pointing out enlarged regional lymph nodes.
Pleural: Span pointing out pleural/pericardial effusion/dissemination.
Distant: Span pointing out distant metastasis outside the lung parenchyma.

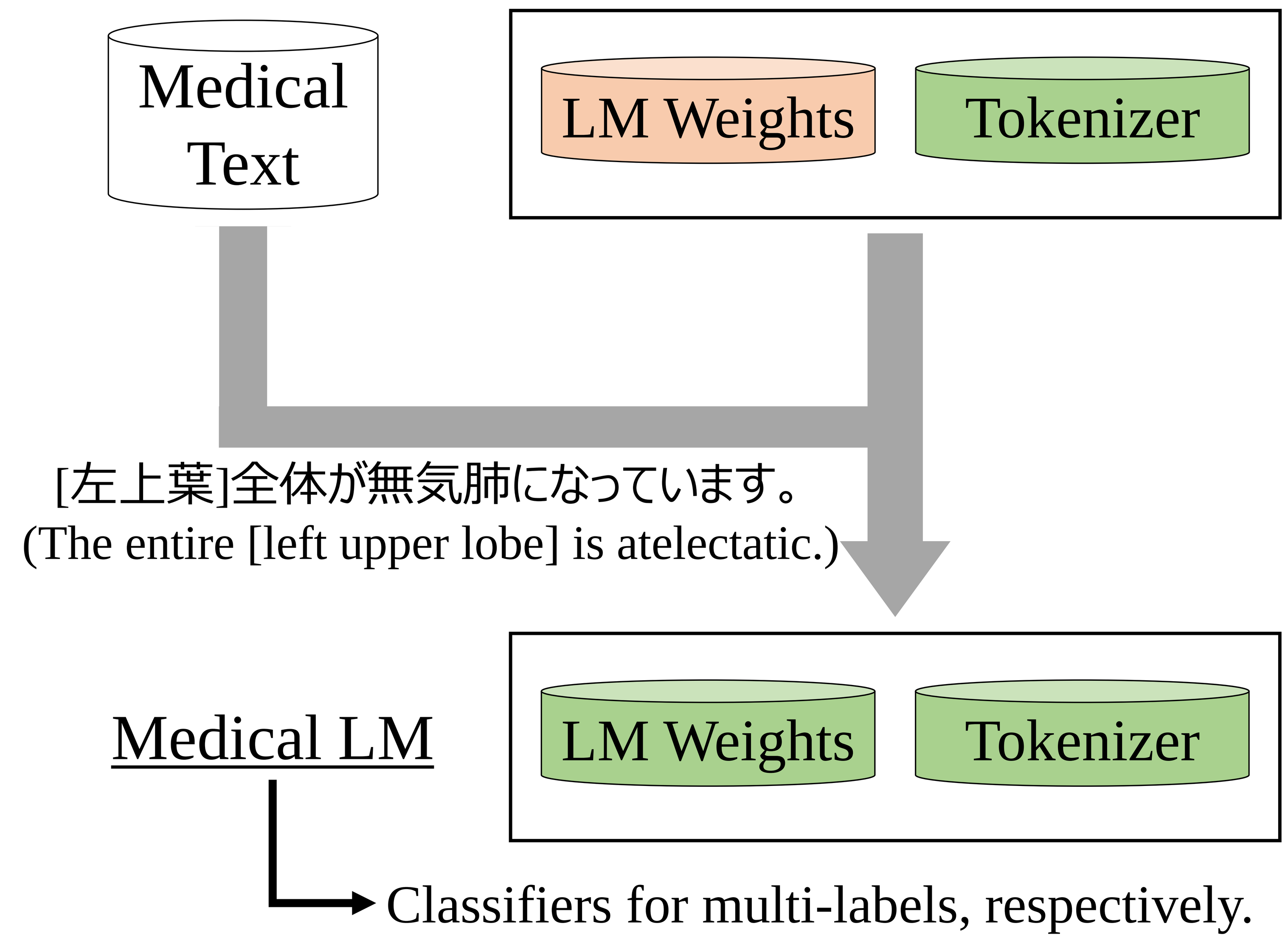
*Image and texts are quoted from the RadNLP sites.

Proposed Method

(1) Tokenizer Adaptation



(2) Continually Pre-Training (CPT)



Results and Discussions

Validation Results on Japanese task

Model	Tokenizer	CPT	Rule※	Avg.
BERT-base	×	×	✓	0.891
	×	✓	×	0.862
	✓	×	×	0.892
	✓	✓	×	0.905
	✓	✓	✓	0.912

※The rules applied here are those defined by the Japanese TNM classification and optimized using validation data.

Validation Results on English task

Model	Tokenizer	CPT	Rule	Avg.
BERT-base	✓	✓	×	0.914
	✓	✓	✓	0.909

Test Results

Language	Tokenizer	CPT	Rule	Avg.
Japanese	×	×	✓	0.824
English	✓	×	×	0.886

※The organizers only showed the best model and its score.

#1 Effects of CPT

- ✓ CPT hardly effects the performance of both Japanese / English sub tasks.

#2 Effects of Tokenizer Update

- ✓ This process is better to adopt for LM, but not necessary when the rules are well-constructed.

#3 Necessity of Data Cleaning

- ✓ There were instances of clearly incorrect labeling (e.g., labels assigned to sentences that do not fit any category).
- ✓ The Japanese dataset included sentences that appeared to be translations from English. resulting in slightly unnatural expressions.
- ✓ Radiology reports typically reflect variations influenced by institutional preferences and educational backgrounds.
  →This kind of information will need to be taken into consideration