Domain Adaptation with Medical Vocabulary-Aware Tokenizer for Radiology Report Analysis in RadNLP at KAIY003 Daiki Shirafuji¹ Takafumi Niwa²

¹Mitsubishi Electric Corporation ²National Defense Medical College

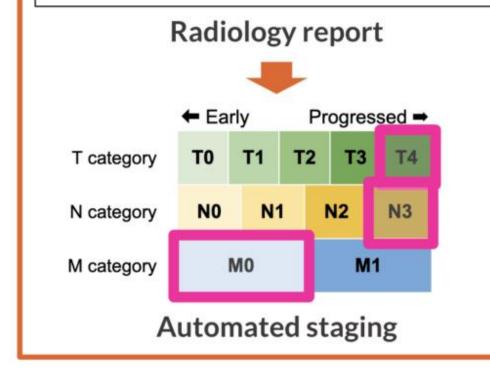
Explanation: Sub Task on RadNLP

Scope of RadNLP 2024



Lung cancer CT image

A tumor with a diameter of 12cm is observed spreading across the upper and lower lobes of the left lung, suggesting known lung cancer. It is in extensive contact with the pleura and is accompanied by the destruction of the left 3rd rib. Rib and parietal pleural infiltration is suspected. There are small nodules in the left upper lobe, suspecting secondary tumor nodules. The left mediastinal and bilateral hilar lymph nodes are enlarged, suspecting metastasis. No pleural effusion is observed. No obvious abnormalities are observed in the upper abdominal organs within the imaging range.



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.

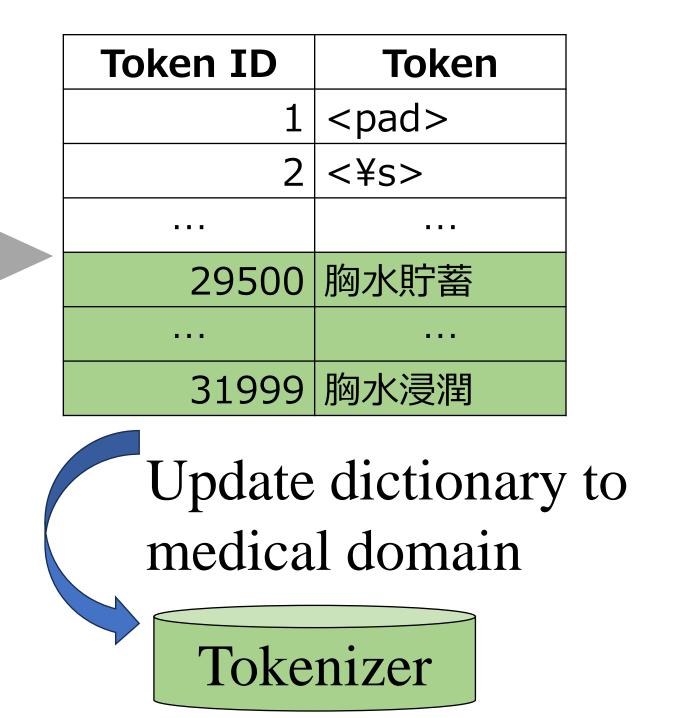
*Image and texts are quoted from the RadNLP sites.

Proposed Method

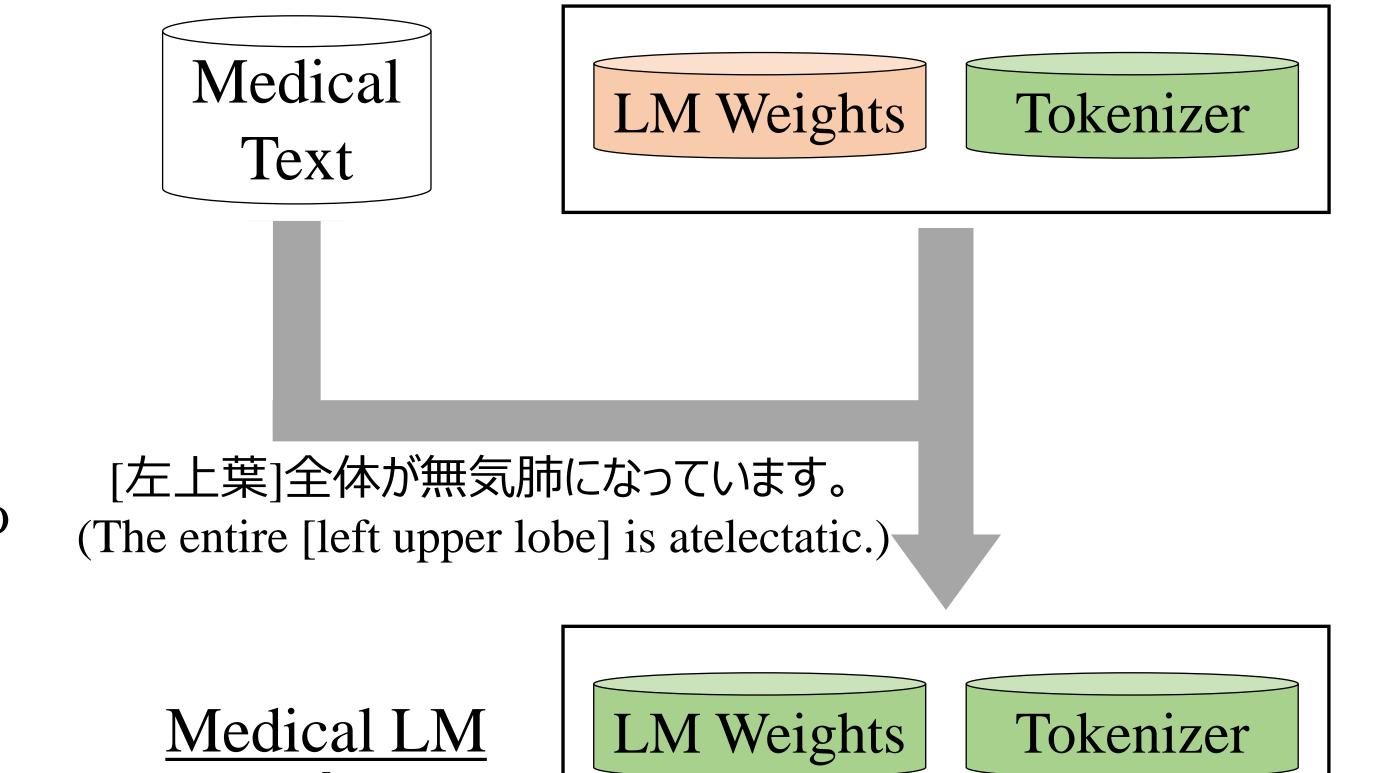
(1) <u>Tokenizer Adaptation</u>

Wiki	40b ——		
	Token ID	Token	Counter
	1	<pad></pad>	50,000
	2	<¥s>	30,000
	• • •	• • •	•••
	29500	周年を迎えた	500
	31999	記録を樹立	30

Token	Counter	
胸水/貯蓄	5,000	_
胸水/浸潤	2,000	



(2) <u>Continually Pre-Training (CPT)</u>





regard bi/tri-gram as one series, and extract top-K frequent tokens

Results and Discussions

Validation Results on Japanese task

Model	Tokenizer	СРТ	Rule🔆	Avg.
BERT-base	X	X	\checkmark	0.891
	X	\checkmark	X	0.862
	\checkmark	X	X	0.892
	\checkmark	\checkmark	X	0.905
	\checkmark	\checkmark	\checkmark	0.912

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

#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

 \checkmark 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

Validation Results on English task

Model	Tokenizer	CPT	Rule	Avg.
BERT-base	\checkmark	\checkmark	X	0.914
	\checkmark	\checkmark	\checkmark	0.909

Test Results

Language	Tokenizer	CPT	Rule	Avg.
Japanese	X	X	\checkmark	0.824
English	\checkmark	X	X	0.886
X The organizers only showed the best model and its score.				