NITKC at the NTCIR-18 RadNLP shared task: Using Graph-RAG in a lung cancer staging method with Natural Language Processing for Radiology

Aoi Kondo*, Tan You Quan Bernon**, Tsubasa Oka*, Hiroaki Koga* and Mikio Oda*
(*NIT, Kurume College, **Temasek Polytechnic)

06-11-2025 / The 18th NTCIR: RadNLP2024 Shared Task / Tokyo, Japan

【Background】

The RadNLP 2024 Shared Task [1]

- The RadNLP 2024 Shared Task of **Natural Language Processing (NLP) for Radiology**.
- Consists of two tasks called the **main task** and the **subtask**.

Main task

- Multi-label **document classification**.
- To predict the **T, N, and M categories (TNM classification)** for each radiology report.
- The TNM classification is a **hierarchical structure.**
  - The T category contains **T0, T2, T3 T4 and Tis, with T1mi, T1a, T1b, T1c, T2a, and T2b below them.**
  - The N category contains **N0, N1, N2, and N3.**
  - The M category contains **M0 and M1, with M1a, M1b, and M1c below M1.**

Subtask

- Multi-label **sentence binary classification**.
- Each sentence is checked for **multiple lung cancer–related topics:** Omittable, Measure, Extension, Atelectasis, Satellite, Lymphadenopathy, Pleural, and Distant.
- The model predicts whether each topic is **mentioned or not (True/False) for each sentence**.

【Proposed Methods】

Main task

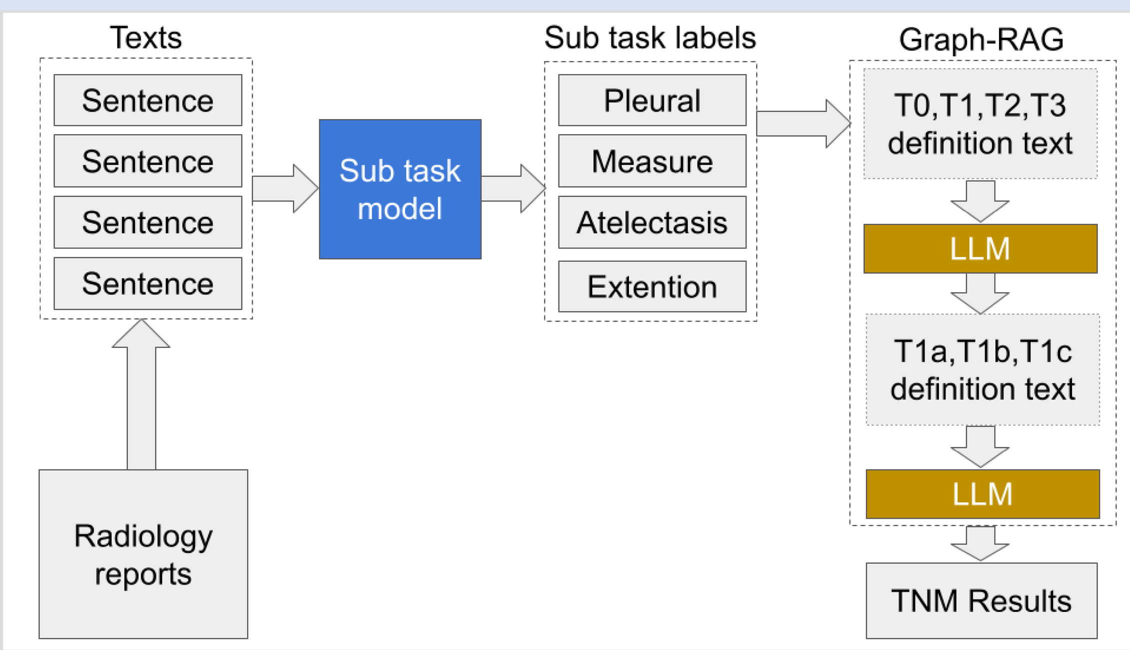
- We adopt the **Graph-RAG** [2] approach **to determine the TNM category.**
- We use the **subtask results**.

Subtask

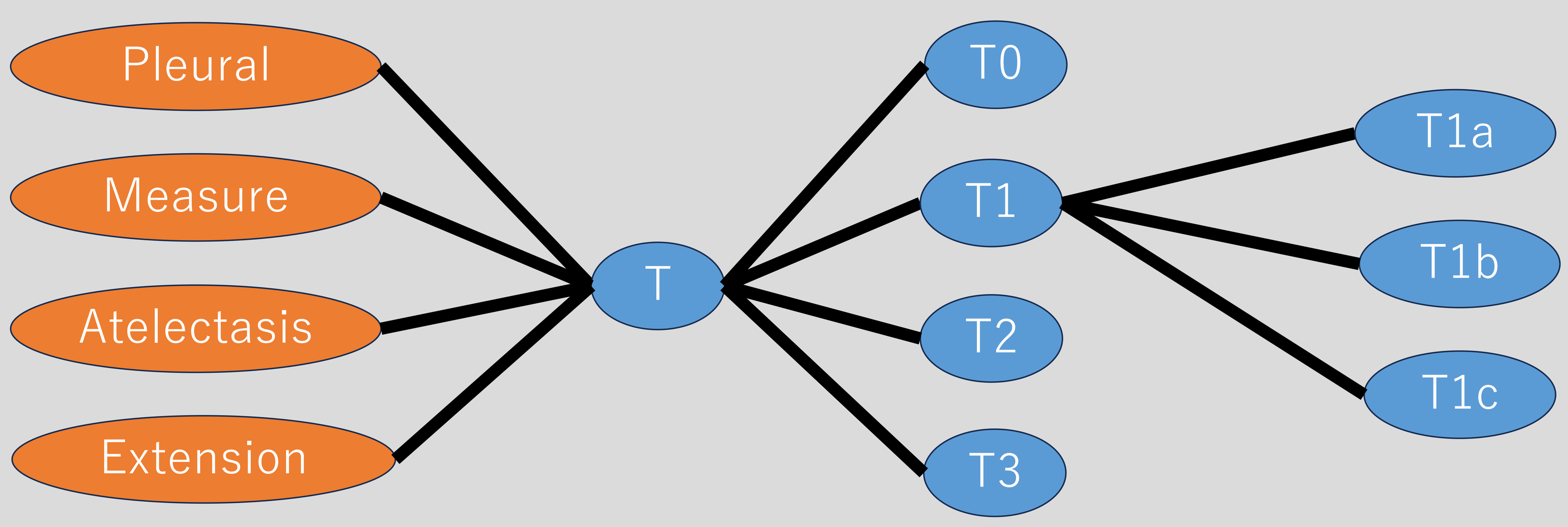
- We use **BioBERT** and **MedBERT** to predict labels.
- BioBERT and MedBERT are pre-trained models for **the medical NLP task**.

How to predict the labels of the main task

1. Divide a radiology **report into sentences**.
2. Perform **binary classification** for each subtask topic.
3. Adapt the definitions of the topics **based on the binary classification**, and use **Graph-RAG** to insert them **into the LLM's prompts**.



An overview of our method for the main task



The structure of the T1 classification

【Experiments】

Experimental setup

- We used **sometimesanotion/Lamarck-14B-v0.7** model as the LLM for the main task.
- We used **Neo4j** for the graph database.
- The subtask was trained for 10 **epochs with a batch size of 4**.

Evaluation methods

- We used two types of evaluation methods: **fine-grained and coarse-grained**.
  - The fine-grained score is the proportion of reports where **all T, N, and M factors are correctly predicted**.
  - The coarse-grained score **ignores the subcategories of the TNM classes**.

Accuracy scores of the main task

Evaluation type	Fine	Coarse
Joint accuracy	0.296	0.482
T accuracy	0.457	0.642
N accuracy	**0.864**	**0.864**
M accuracy	0.778	0.815

【Discussions】

Are our methods really effective?

- We compared **our method** with **the Long-Context (LC)** approach using validation data from **the main task**.
  - The Long-Context method is a **data augmentation approach** that uses an LLM prompt to insert all definition text **directly**.
  - In contrast, our method inserts definition **text based on a graph structure**.
- Our method **outperformed** LC in both fine-grained and coarse-grained evaluations.

Comparison between our method and LC

	Our method		Long-Context	
	Fine	Coarse	Fine	Coarse
Joint accuracy	0.500	0.667	0.273	0.527
T accuracy	0.611	0.796	0.473	0.746
N accuracy	**0.907**	**0.907**	0.764	0.764
M accuracy	0.852	0.889	**0.782**	**0.837**

【Conclusions】

- We used **Graph-RAG** for the main task and **BERT** models for the subtask.
- In future work, we plan to **enhance the graph** with domain-specific medical knowledge in the main task and to train the subtask model with a **larger dataset**.

【References】

[1] Yuta Nakamura et al., “NTCIR-18 RadNLP 2024 Overview: Dataset and Solutions for Automated Lung Cancer Staging,” in In Proceedings of the NTCIR-18 Conference, June 2025.

[2] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” Neural Inf Process Syst, vol. abs/2005.11401, pp. 9459–9474, May 2020.