TUSNLP at the NTCIR-18 RadNLP Task:
Explainable Classification Approach by Domain Knowledge-Based Bag-of-Words

Tomoki Terada[1] and Rei Noguchi[1*]

[1]Department of Business Economics, School of Management, Tokyo University of Science

*Corresponding author.

Introduction

Radiology reports contain important findings provided by radiologists; however, since the final diagnosis is made by attending physicians, critical information can be overlooked in the busy clinical field. Therefore, automatically extracting key findings from radiology reports is valuable for supporting diagnosis. In this study, we aim to develop an interpretable machine learning model that determines the stage of lung cancer from radiology reports.

Method

Data Processing

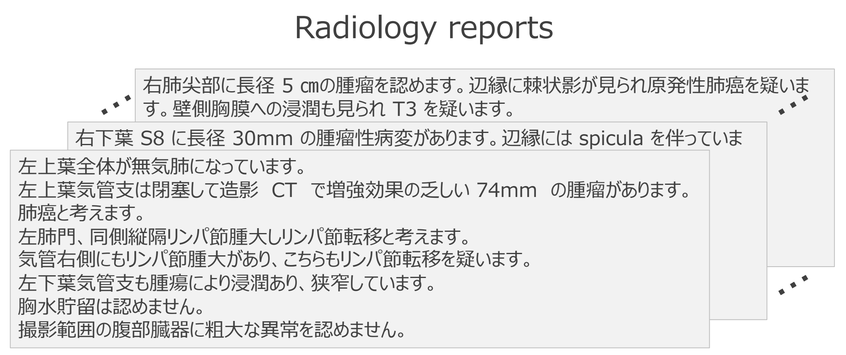
- Defined key terms for each stage category based on lung cancer staging criteria published by the Japan Lung Cancer Society (JLCS) and calculated their frequencies of occurrence in the radiology reports. Tumor size information was also extracted directly from the reports.
- Structured the reports into a word frequency table (a kind of Bag-of-Words), with additional processing such as negation detection and selective sentence filtering to improve feature quality.

Model Development

- Developed classification models to determine the T, N, and M stages of lung cancer from the reports, along with a regression model to estimate tumor size, which is an important feature for T staging.
- Random Forest, LightGBM, and CatBoost were utilized for classification and regression tasks, with model performance enhanced through feature selection, class balancing (using SMOTE), hyperparameter tuning, and ensemble methods.

Model Evaluation

- Evaluated the models using stratified train-test splits and standard metrics such as precision, recall, and (weighted) F1-score, with feature importance visualization to enhance model interpretability for clinical use.



Report ID	T/N/M stages of lung cancer			腫瘤	肺癌	胸部異常陰影	肝	腫瘤(⊘)	浸潤	内部	癌	…
	T	N	M									
56344	T4	N3	M0	0	1	0	0	0	0	1	0	
13316	T1c	N0	M0	1	1	1	1	0	1	0	0	
16574	T3	N0	M0	1	1	0	1	1	0	0	1	
40488	T4	N0	M1c	1	0	1	0	0	1	1	1	
:			:	:								

Figure 1 Schematic view of word frequency table

Results

- For training and validation data, the fine-grained accuracies for T, N, and M classifications exceeded 94%, with a Joint accuracy of 90.7%, showing high performance across all models.
- Feature importance analysis demonstrated that medically relevant features, such as tumor size for T classification and lymph-related terms for N classification, significantly contributed to prediction accuracy.
- Notably, terms related to metastasis and specific organs (e.g., kidney, bilateral findings) also influenced M classification, indicating that the model captured complex relationships beyond predefined criteria.
- However, in the formal test evaluation, classification accuracy, especially for T classification, significantly dropped, suggesting overfitting and the need for improved handling of tumor size estimation and localization features.

Table 1 Training, validation, and test (formal run) results

	Fine				Coarse			
	Joint accuracy	T accuracy	N accuracy	M accuracy	Joint accuracy	T accuracy	N accuracy	M accuracy
Train	0.9074	0.9630	0.9537	0.9815	0.9074	0.9630	0.9537	0.9815
Validation	0.9074	0.9630	0.9444	0.9815	0.9259	0.9815	0.9444	0.9815
Test (formal run)	0.2176	0.3519	0.8287	0.7963	0.3796	0.5000	0.8287	0.8611

Table 2 to 4 Feature importances in T, N and M classifications, respectively

	Feature	Importance	Feature	Importance	Feature	Importance
1	max_mm	0.043783	リンパ節_frequency3	0.047007	M1cキーワード合計2	0.026812
2	左_frequency1	0.021610	リンパ節_frequency2	0.037994	腎_frequency3	0.025907
3	左_frequency2	0.021019	リンパ_frequency2	0.036443	転移_frequency2	0.022088
4	縦隔_frequency2	0.020570	リンパ_frequency3	0.030953	腎_frequency1	0.019802
5	浸潤_frequency1	0.019226	腫大_frequency3	0.024422	M1cキーワード合計1	0.018964
6	浸潤_frequency2	0.018517	N2キーワード合計	0.023184	腎_frequency2	0.018638
7	転移_frequency2	0.011432	転移_frequency2	0.022436	多発_frequency2	0.017365
8	リンパ節_frequency2	0.009752	縦隔_frequency2	0.021020	多発_frequency1	0.016124
9	N0	0.009691	転移_frequency1	0.018997	転移_frequency3	0.015139
10	腫瘤_frequency1	0.009628	腫大_frequency2	0.017710	両側_frequency2	0.012323

Conclusion

- Highly interpretable classification models were successfully developed by predefining key terms based on domain knowledge, such as clinical guidelines, and by using their frequencies as training data. The models had high medical validity and provided new insights, such as the contribution of the keyword “kidney” in the M classification model.
- This method is versatile and likely equally applicable to any disease for which guidelines are available.
- On the other hand, there are some limitations, and if these are resolved, the method becomes even more useful and valuable:
  - Solve the problem of overfitting and improve model generalization performance.
  - Automate the pre-definition of key terms by analyzing the guidelines textually.