NTCIR 17 Conference: Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan
DOI: https://doi.org/10.20736/0002001285

NAISTSOCRR at the NTCIR-17 MedNLP-SC Radiology Report Subtask

Takuya Fukushima
Nara Institute of Science and Technology
Japan
fukushima.takuya.fo2@naist.ac.jp

Yuka Otsuki
Nara Institute of Science and Technology
Japan
otsuki.yuka.pb0@naist.ac.jp

Shuntaro Yada
Nara Institute of Science and Technology
Japan
s-yada@naist.ac.jp

Shoko Wakamiya
Nara Institute of Science and Technology
Japan
wakamiya@naist.ac.jp

Eiji Aramaki
Nara Institute of Science and Technology
Japan
aramaki@naist.ac.jp

ABSTRACT

This paper describes how we tackled the Medical Natural Language Processing for Radiology Report TNM staging (RR-TNM) Subtask as participants of NTCIR17. The RR-TNM Subtask is a MedNLP-SC original task to classify radiology reports under multiple criteria. We introduced three different methods based on pre-trained language models (PLMs), including a medical-specific model. Notably, our combination approach, utilizing JMedRoBERTa (manbyo-wordpiece) for label T, Tohoku-BERT-v3 for label N, and UTH-BERT for label M, achieved an accuracy of 0.3704 on the test data. This performance was the highest among all participants, emphasizing the effectiveness of our strategy.

KEYWORDS

Medical Natural Language Processing, Case Reports, Radiography Reports, Case Identification

TEAM NAME

NAISTSOCRR

SUBTASKS

RR-TNM

1 INTRODUCTION

Recently, the shift from traditional paper-based medical records to electronic format has become more pronounced, making the accessibility and storage of patient data more streamlined. This digital transformation not only modernizes healthcare documentation but also underscores the ever-growing significance of information processing techniques in medical fields. The capability to swiftly and accurately process vast amounts of data can greatly influence diagnosis, treatment decisions, and research outcomes.

In such a situation, we tackled MedNLP-SC Subtask 2: Radiology Report TNM staging (RR-TNM). This task is to classify radiology reports based on the TNM system, a globally recognized framework for categorizing cancer severity and spread. This system consists of three key dimensions: tumor (T), lymph node (N), and metastasis (M), effectively translating into three simultaneous multiclass classification tasks. Labels range from early stages like T1N0M0 to more advanced ones such as T3N2M1. While TNM staging is essential for treatment and research, it traditionally requires intricate reasoning by medical professionals. An inherent challenge in this realm is the periodic "stage migration" due to the revision of cancer staging criteria. The necessity to reclassify under updated guidelines, especially in large patient databases, underscores the potential value of efficient NLP systems in medical advancements. Further details of this task can be found in the NTCIR MedNLP-SC website[1] and the NTCIR-17 MedNLP-SC task overview paper [3].

We implemented three standard methods based on several pre-trained language models (PLMs). As the PLMs, we used the representative Japanese models including those trained on medical datasets. The three proposed methods have distinct approaches to handling the training objectives. The first method employs whole TNM label combinations as distinct labels. The second method undertakes multi-task learning using three separate labels (T, N, and M) as objective variables. to predict three labels separately as multi-label classification.The third method trains a separate model for each of the three labels T, N, and M. The inherent advantage of our method's straightforwardness is its reproducibility. It does not only demonstrate the potential of NLP in the medical field, but also paves the way for future studies by providing a replicable foundation.

2 MATERIALS AND METHODS

The RR-TNM dataset consists of 253 Japanese radiology reports originating from the open access image data sets. Thus, it can be distributed without privacy concerns. These radiology reports were written by nine certified radiologists, who diagnosed the same open 27 lung cancer cases independently. The 27 cases was split into the training set (108 documents for 12 cases), validation set (54 documents for six cases), and test set (81 documents for nine cases).

We introduce our methods consisting of a baseline and the proposed three PLM-based methods.

2.1 Majority baseline

The majority baseline always return the most frequent TNM label (i.e., T1N0M0) as the prediction result. The top 5 frequently labels

[1] https://sociocom.naist.jp/mednlp-sc/

in the train data are as follows: T1N0M0 (15 instances), T2N0M0 (13 instances), T4N2M0 (13 instances), T2N2M1 (11 instances) and T3N1M1 (7 instances).

2.2 PLM-based methods

We propose three methods based on pre-trained language models (PLMs). We adopted BERT, a Transformers-based large-scale language model [1], because of its representativeness of the PLMs. To solve RR Subtask, we designed three training objectives:

Unilabel single-model Unilabel single-model fine-tuned by single label, which is a joint of three labeled of T, N, and M.

Multilabel single-model Multilabel single-model treating T, N, and M independently as three distinct labels. We employed a multi-task learning paradigm for fine-tuning the model.

Multilabel multi-models Multilabel multi-models employs a segmented approach by independently fine-tuning three distinct models, each with T, N, and M as their individual training objectives.

For these objectives, we fine-tuned the following pre-trained BERT models:

Tohoku-BERT-v2[2]
The Tohoku-BERT-v2 is the commonly used Japanese pre-trained language model. This model was trained on the Japanese version of Wikipedia. The training corpus is sourced from the Wikipedia Cirrussearch dump file dated August 31, 2020.

UTH-BERT [2]
The UTH-BERT is a Japanese model trained on medical data. Specifically, it was pre-trained using diagnostic records.

JMedRoBERTa (sentencepiece) [4]
The JMedRoBERTa (sentencepiece) is a Japanese model trained on medical data, specifically pre-trained using medical articles. It employs the SentencePiece tokenizer, which directly tokenizes sentences into subwords.

JMedRoBERTa (manbyo-wordpiece) [4]
The JMedRoBERTa (manbyo-wordpiece) utilizes the "Manbyo" dictionary, a large-scale disease name dictionary, in conjunction with Mecab for tokenization. It then employs the WordPiece method to further tokenize into subwords.

Tohoku-BERT-v3[3]
The Tohoku-BERT-v3 is enhanced versions of the Tohoku-BERT-v2, having been trained on a more extensive dataset. These models were trained on both the Japanese segment of the CC-100 dataset and the Japanese version of Wikipedia. For the latter, a text corpus was generated from the Wikipedia Cirrussearch dump file as of January 2, 2023.

Tohoku-BERT-v3-char[4]
Tohoku-BERT-v3-char is a varidation of Toholu-BERT-v3 that employs character-level tokenization.

We set the hyperparameters for fine tuning to the values below:

- Batch size: 16
- Learning rate: 2e-5
- Max length: 512
- Early stopping rounds: 5
- Criterion: Cross Entropy Loss
- Optimization: AdamW

[2]https://huggingface.co/cl-tohoku/bert-base-japanese-v2
[3]https://huggingface.co/cl-tohoku/bert-base-japanese-v3
[4]https://huggingface.co/cl-tohoku/bert-base-japanese-char-v3

3 RESULTS

Table 1 shows the validation results of the Unilabel single-model performance. The results demonstrated that JMedRoBERTa (manbyo-wordpiece) achieved the best performance with macro F1 of 0.1510 and Accuracy of 0.4259.

Table 2 shows the validation results of the Multilabel single-model performance. The results demonstrated that JMedRoBERTa (sentencepiece) recorded the best performance with macro F1 of 0.2147 and Accuracy of 0.4630.

Table 3 shows the validation results of the Multilabel multi-models performance. The results demonstrated that JMedRoBERTa (manbyo-wordpiece) recorded the best performance with macro F1 of 0.2278 and Accuracy of 0.5000. Finally, combining the prediction results of the following three models, which had the best prediction performance for T, N, and M. We obtain macro F1 of 0.2932 and Accuracy of 0.5926.

- T: JMedRoBERTa (manbyo-wordpiece)
- N: Tohoku-BERT-v3
- M: UTH-BERT

We submitted three prediction results for the test data using Baseline, JMedRoBERTa (sentencepiece) on the Multilabel single-model, and Combination (the best models for T, N, and M) on the Multilabel multi-model. Table 4 shows the accuracy in the test data. Similar to the results for the validation data, Combination (the best model for T, N, and M) on the Multilabel multi-models was the best not only among our methods but also against all other participants.

4 DISCUSSION

In order to accurately predict the TNM staging, the prediction performance of T, N, and M all need to be improved. In particular, compared to M, T and N still have challenges in prediction performance. In this paper, we introduced a PLM-based method, but it is worthwhile to try to take advantage of recent methods based on large language models. This is because insufficient number of data is a problem in the medical field, including the present task. Large language models such as GPT-4 are expected to be utilized because of their high performance in Few-shot learning.

To provide clarity on prediction inaccuracies, we conducted an error analysis using confusion matrices for each label: T, N, and M. Figure 1 shows the confusion matrix in the validation data.

From the analysis, the major sources of errors for label T, the primary source of error was misclassification between classes 1 and 3. Label N mainly struggled with distinguishing between classes 0 and 2 and also between classes 3 and 2. For label M, there was a balanced misclassification between the two classes. Based on the error analysis, Future efforts should focus on enhancing the model's discernment between closely related classes in labels T and N, and addressing the balanced errors in label M.

Table 1: Validation score of Unilabel single-model

	T		N		M		All	
Model Name	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1	Accuracy
Baseline (most frequency label)	0.1250	0.3333	0.1667	0.5000	0.3333	0.5000	0.0500	0.3333
Tohoku-BERT-v2	0.4292	0.5556	0.3744	0.7037	0.6589	0.6852	0.1093	0.3333
UTH-BERT	0.4461	**0.6111**	0.4223	**0.7778**	0.6506	0.6852	0.1307	0.3704
JMedRoBERTa (sentencepiece)	0.3967	0.4074	**0.4520**	0.7037	0.7875	**0.7963**	0.0923	0.2593
JMedRoBERTa (manbyo-wordpiece)	**0.4563**	0.5741	0.4201	0.7593	**0.7905**	**0.7963**	**0.1510**	**0.4259**
Tohoku-BERT-v3	0.4109	0.5185	0.3858	0.7037	0.7875	**0.7963**	0.1371	0.3519
Tohoku-BERT-v3-char	0.3911	0.5185	0.4137	0.7593	0.7444	0.7593	0.1251	0.3519

* Bold font indicates the best score for each evaluation metric.

Table 2: Validation score of Multilabel single-model

	T		N		M		All	
Model Name	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1	Accuracy
Baseline (most frequency label)	0.1250	0.3333	0.1667	0.5000	0.3333	0.5000	0.0500	0.3333
Tohoku-BERT-v2	0.4001	0.4444	0.3964	0.7037	0.7699	0.7778	0.1757	0.4444
UTH-BERT	0.5018	0.4815	0.3933	0.6852	0.8286	**0.8333**	0.1367	0.3333
JMedRoBERTa (sentencepiece)	0.6202	0.5926	**0.5566**	0.7593	0.8107	0.8148	**0.2147**	**0.4630**
JMedRoBERTa (manbyo-wordpiece)	**0.6436**	**0.6852**	0.4223	**0.7778**	0.8107	0.8148	0.2085	0.4444
Tohoku-BERT-v3	0.3866	0.3889	0.4084	0.7037	**0.8319**	**0.8333**	0.1403	0.3519
Tohoku-BERT-v3-char	0.3602	0.4630	0.4268	**0.7778**	0.7444	0.7593	0.1148	0.3148

* Bold font indicates the best score for each evaluation metric.

Table 3: Validation score of Multilabel multi-models

	T		N		M		All	
Model Name	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1	Accuracy
Baseline (most frequency label)	0.1250	0.3333	0.1667	0.5000	0.3333	0.5000	0.0500	0.3333
Tohoku-BERT-v2	0.4764	0.4815	0.6228	0.8148	0.8082	0.8148	0.1274	0.3704
UTH-BERT	0.5999	0.5741	0.5600	0.7407	**0.9259**	**0.9259**	0.1570	0.4074
JMedRoBERTa (sentencepiece)	0.5778	0.5926	0.6962	0.7778	0.8875	0.8889	0.1863	0.4630
JMedRoBERTa (manbyo-wordpiece)	**0.6668**	**0.7407**	0.6759	0.7963	0.8107	0.8148	**0.2278**	**0.5000**
Tohoku-BERT-v3	0.3471	0.3704	**0.7023**	**0.8333**	0.8107	0.8148	0.0537	0.2593
Tohoku-BERT-v3-char	0.4646	0.4815	0.4315	0.7963	0.8692	0.8704	0.1673	0.3704
Combination (best model for T, N, and M)	**0.6668**	**0.7407**	**0.7023**	**0.8333**	**0.9259**	**0.9259**	**0.2932**	**0.5926**

* Bold font indicates the best score for each evaluation metric.

5 CONCLUSION

In this paper, we presented the methods of the NAISTSOCRR team in the NTCIR17 MedNLP-SC Radiology Report (RR) Task. We employed three methods based on pre-trained language models, including those trained on medical data, in addition to a baseline with most frequent labels. As a result, the best-performing models differed for each of T, N, and M, and the best predictions for TNM staging were achieved by combining the best models from each label. This result suggested that the independent label handling shows the feasibility against the limited data task.

ACKNOWLEDGEMENTS

This work was supported by JST, AIP Trilateral AI Research, Grant Number JPMJCR20G9, JST AIP-PRISM Grant Number JPMJCR18Y1, Japan, as well as ANR grant ANR-20-IADJ-0005-01, France, under the trilateral ANR-DFG-JST AI call.

Table 4: Accuracy in test data

Method	T	N	M	All
Baseline (most frequency label)	0.3086	0.4198	0.5926	0.1975
JMedRoBERTa (sentencepiece) on the Multilabel single-model	0.6049	0.6049	0.8765	0.3086
Combination on the Multilabel multi-models	0.6667	0.5679	0.8395	**0.3704**
Team A's best model	0.4815	0.6049	0.7407	0.2346
Team B's best model	0.3951	0.8025	0.9259	0.2716

* Bold font indicates the best score for each evaluation metric.

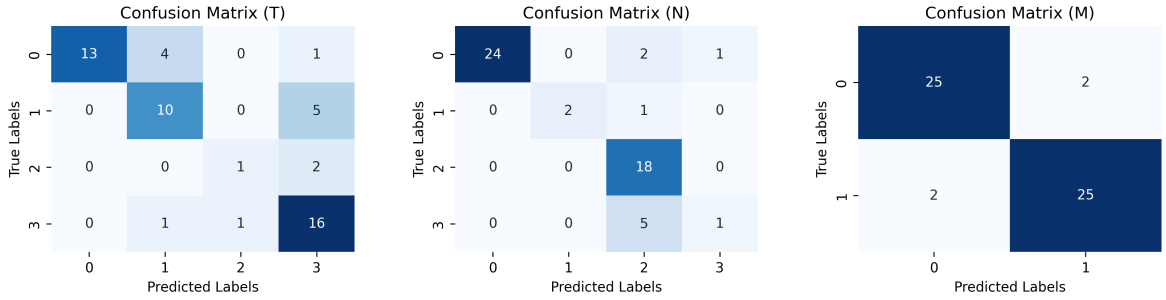

Figure 1: Confusion matrix in validation data

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[2] Yoshimasa Kawazoe, Daisaku Shibata, Emiko Shinohara, Eiji Aramaki, and Kazuhiko Ohe. 2021. A clinical specific BERT developed using a huge Japanese clinical text corpus. *PLoS One* 16, 11 (Nov. 2021), e0259763.

[3] Yuta Nakamura, Shouhei Hanaoka, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2023. NTCIR-17 MedNLP-SC Radiology Report Subtask Overview: Dataset and Solutions for Automated Lung Cancer Staging. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies*. https://doi.org/10.20736/0002001328

[4] Kaito Sugimoto, Taichi Iki, Yuki Chida, Teruhito Kanazawa, and Akiko Aizawa. 2023. JMedRoBERTa: a Japanese Pre-trained Language Model on Academic Articles in Medical Sciences (in Japanese). In *Proceedings of the 29th Annual Meeting of the Association for Natural Language Processing*. https://www.anlp.jp/proceedings/annual_meeting/2023/pdf_dir/P3-1.pdf