ATILF at NTCIR-18 RadNLP 2024 Shared Task:
With less radiology reports, comes less performance

Aman Sinha[1,2] and Ioana Buhnila[1]

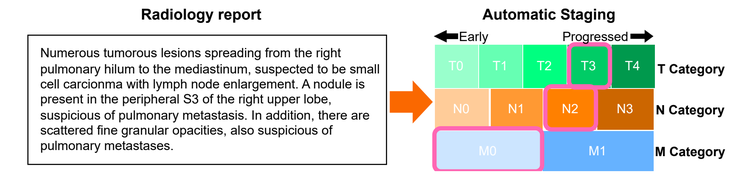
[1]ATILF, Université de Lorraine - CNRS, Nancy, France [2]ICANS Strasbourg
firstname.lastname@univ-lorraine.fr

Introduction

☐ In this work we investigated the performance of general and medical PLMs and LLMs on radiology report identification and classification in English.
☐ We benchmarked PLMs and LLMs for TNM clinical staging classification and sentence segmentation classification task using prompts and class/subclass definitions to perform better semantic disambiguation.
☐ Our results showed that in low amount of data setting, we can obtain better results with medical PLMs in comparison to general and medical LLMs.

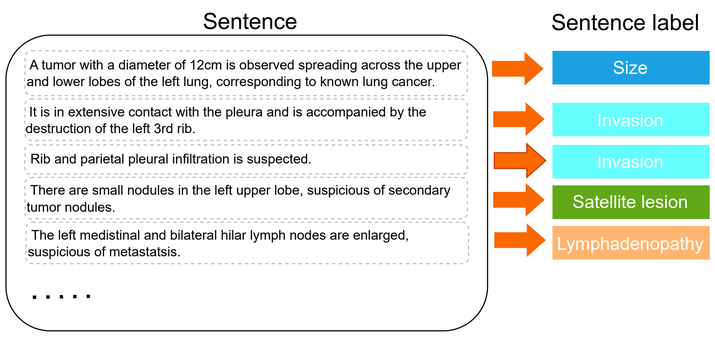
Task Description

► Main Task: TNM clinical staging classification.



Number of classes : T Category (10), N Category (4) and M Category (4).
Total number of possible classes : 160.

► Sub Task: Sentence Segmentation Classification.



Number of classes: Inclusion(2), Measure(2), Extension(2), Atelectasis(2), Satellite(2), Lymphadenopathy(2), Pleural(2), Distant(2).
Total number of possible classes: 64.

Experimental Protocol - Main Task

► TNM clinical staging classification.

☐ **Baseline PLMs:** BioBERT and BioClinicalBERT.

☐ **Zero-Shot Prompting:** (i) Simple-Prompting(P) (ii) Definition-Prompting(D);
LLMs: Llama3.2 , BioMedllama

► **Evaluation metrics:** Individual Accuracy and Joint Accuracy (Fine and Coarse).

Experimental Protocol - Sub Task

► Sentence segmentation classification.

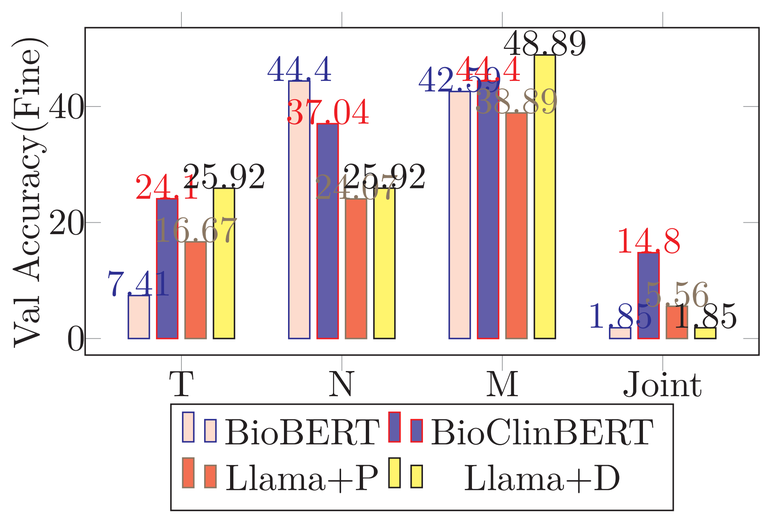
☐ **Baselines:** Unsupervised Clustering.

☐ **PLMs:** BioBERT, BioClinicalBERT, and ClinicalBigBIRD.

► **Loss Function:** Focal Loss

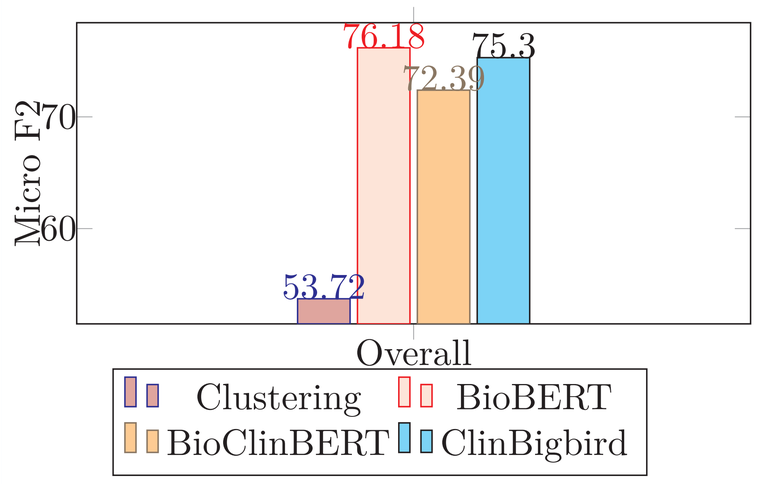
► **Evaluation metrics:** F2.0 Micro Scores.

Val Results - Main Task



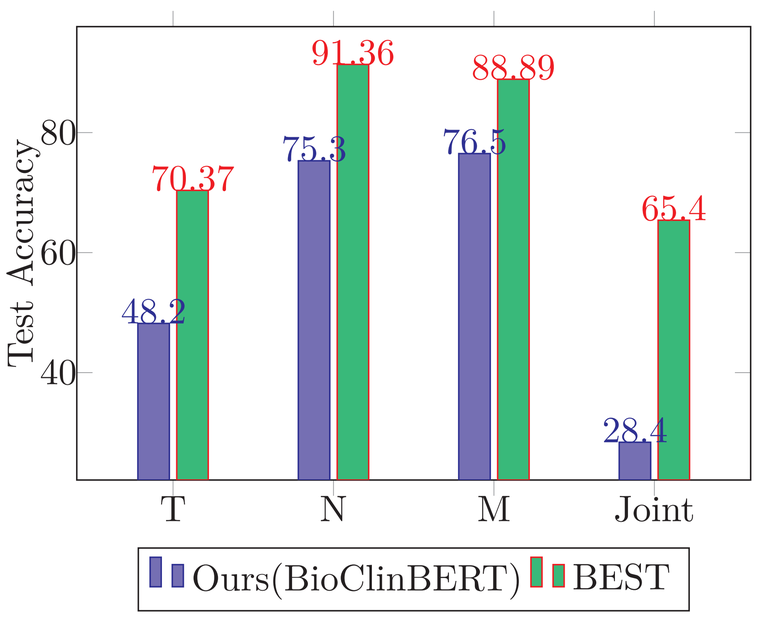
► **Our final system:** BioClinicalBERT

Val Results - Sub Task

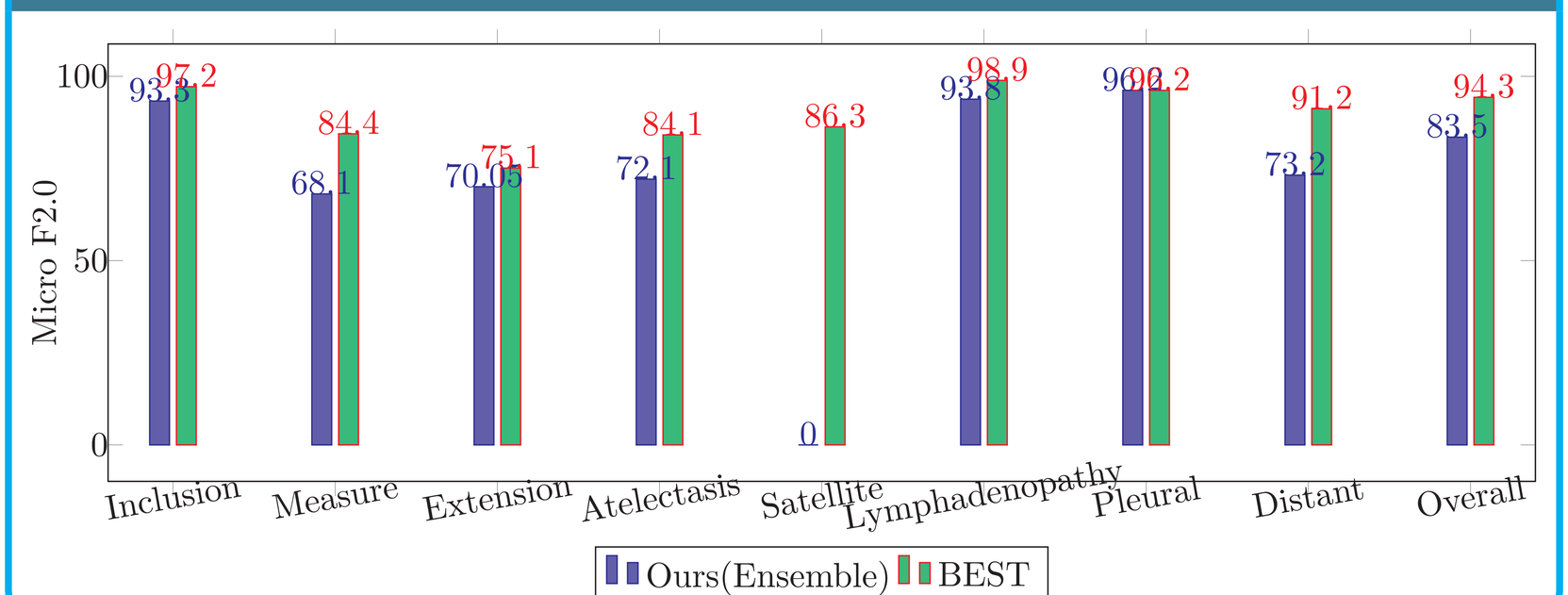


► **Our final system:** Ensemble of BioBERT, BioClinicalBERT, ClinicalBigBIRD.

Test Results - Main Task



Test Results - Sub Task



Findings and Conclusions

► Findings
☐ For clinical staging, PLMs performs competitive compared to zero-shot LLMs.
☐ Definition based prompting is more effective compared to vanilla prompting for individual T/N/M clinical staging identification.
☐ Increasing number of classes M<=N<T lowers LMs capability to automatic clinical staging. Furthermore, **joint accuracy drops strongly for all models due large label space**.
☐ For sentence segmentation, ensemble of PLMs stand very competitive against best systems.

► Conclusions
☐ Automatic clinical staging remains challenging for both small and large language models (the most difficult: T class). However, **PLMs show have advantage over LLMs** as *LLMs suffer from hallucination.*
☐ **PLMs show strong potential** for sentence segmentation classification with data augmentation and further hyper parameter tuning.

Our Paper