An NLP Method to Identify Macro Guideline Sentences in Radiology Reports

Zhaoyi Sun, MS[1], Namu Park, MS[1], Özlem Uzuner, PhD[2], Martin Gunn, MBChB[1], Meliha Yetisgen, PhD[1]

[1]University of Washington, Seattle, WA, USA; [2]George Mason University, Fairfax, VA, USA

Introduction

Radiology reports document imaging study results, highlighting abnormalities and providing diagnoses or treatment recommendations to guide patient care. These reports typically include patient health information alongside system-generated macro guidelines, which offer general recommendations or standardized templates to assist in the interpretation of patient-specific clinical findings. However, macro guidelines can refer to the same clinical concepts that are of interest in patient care, creating a challenge for the evaluation of information extraction (IE) [1] methods that aim to determine the key information related to patient care. If macro guidelines are not filtered, IE models may misinterpret general recommendations as patient-specific findings. Figure 1 shows an example that a size mention from pancreatic cyst management guidelines could be mistakenly tied to a specific patient's cyst, leading to inaccurate extractions and associations. To address this, we present a BERT-based approach using ClinicalBERT[2] to automatically identify and filter macro guidelines in radiology reports, ensuring a clearer focus on patient-specific clinical information.

> Pancreatic cyst < 2 cm in size in an asymptomatic patient with no suspicious imaging features, consider follow-up MRI in one year (if stable after 1 year no further follow-up is needed).

Figure 1. Example macro guideline from a radiology report.

Methods

We treated macro guideline identification as a sentence classification task and used NLTK for sentence tokenization in radiology reports. A gold-standard dataset consisting of 150 radiology reports and 8,860 sentences was annotated by three radiologists. Of these, a senior board-certified radiologist annotated 90 reports, while two research fellows annotated 60 reports. A total of 777 sentences containing macros were labeled as "1", while the remaining 8,083 sentences were labeled as "0". Most macros spanned multiple sentences, with a median length of 5 sentences, and the most common lengths being 4 and 5. This highlights the importance of modeling macro guidelines as sentence sequences rather than isolated sentences. The dataset was split at the report level using 5-fold cross-validation with an 80/20 training-test split to ensure robust evaluation. We used TF-IDF vectorization to represent text features and a linear support vector machine (SVM) for classification as a baseline. We fine-tuned ClinicalBERT, adding a dropout layer and a classifier layer, optimizing with Adam and using positive weighting to handle class imbalance. The model was trained for 10 epochs with a batch size of 32 and a learning rate of 1e-5. We used precision, recall, and F1-score to evaluate performance. To capture sequential dependencies in macros, each input included the previous and current sentences.

Results

The results were aggregated across all test folds. Fine-tuned ClinicalBERT outperformed the baseline TF-IDF + SVM. In the macro guideline category, recall increased from 0.9434 to 0.9730, and F1-score improved from 0.9544 to 0.9705, highlighting ClinicalBERT's superior ability to distinguish macros in radiology reports.

Table 1. Classification performance of the models.

Method	Category	P	R	F
TF-IDF + SVM	note sentence	0.9946	0.9968	0.9957
	macro guideline	0.9657	0.9434	0.9544
ClinicalBERT	note sentence	0.9974	0.9969	0.9972
	macro guideline	0.9680	**0.9730**	**0.9705**

Discussion and Conclusions

This study demonstrates the effectiveness of ClinicalBERT in identifying macros in radiology reports. Removing macros before applying IE methods can enhance dataset quality and reduce noise in downstream tasks. However, certain misclassifications highlight areas for improvement. False negatives often involved macro-introductory captions (e.g., "Incidental findings and recommendations:") and single-sentence macro recommendations, while false positives included non-macro recommendation statements (e.g., "Recommend follow-up imaging."). Refining sequence modeling and incorporating contextual cues could further improve accuracy.

References

1. Reichenpfader D, Müller H, Denecke K. A scoping review of large language model based approaches for information extraction from radiology reports. NPJ Digit Med. 2024 24 Aug;7(1):222.
2. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. arXiv [csCL]. 2019.