Identifying Significance of Discrepancies in Radiology Reports

Arman Cohan, Luca Soldaini, Nazli Goharian

Allan Fong, Ross Filice and Raj Ratwani

SIAM International Conference on Data Mining (SDM 2016)

Workshop on Data Mining for Medicine and Healthcare

1789 GEORGETOWN UNIVERSITY

ir LAB GEORGETOWN UNIVERSITY

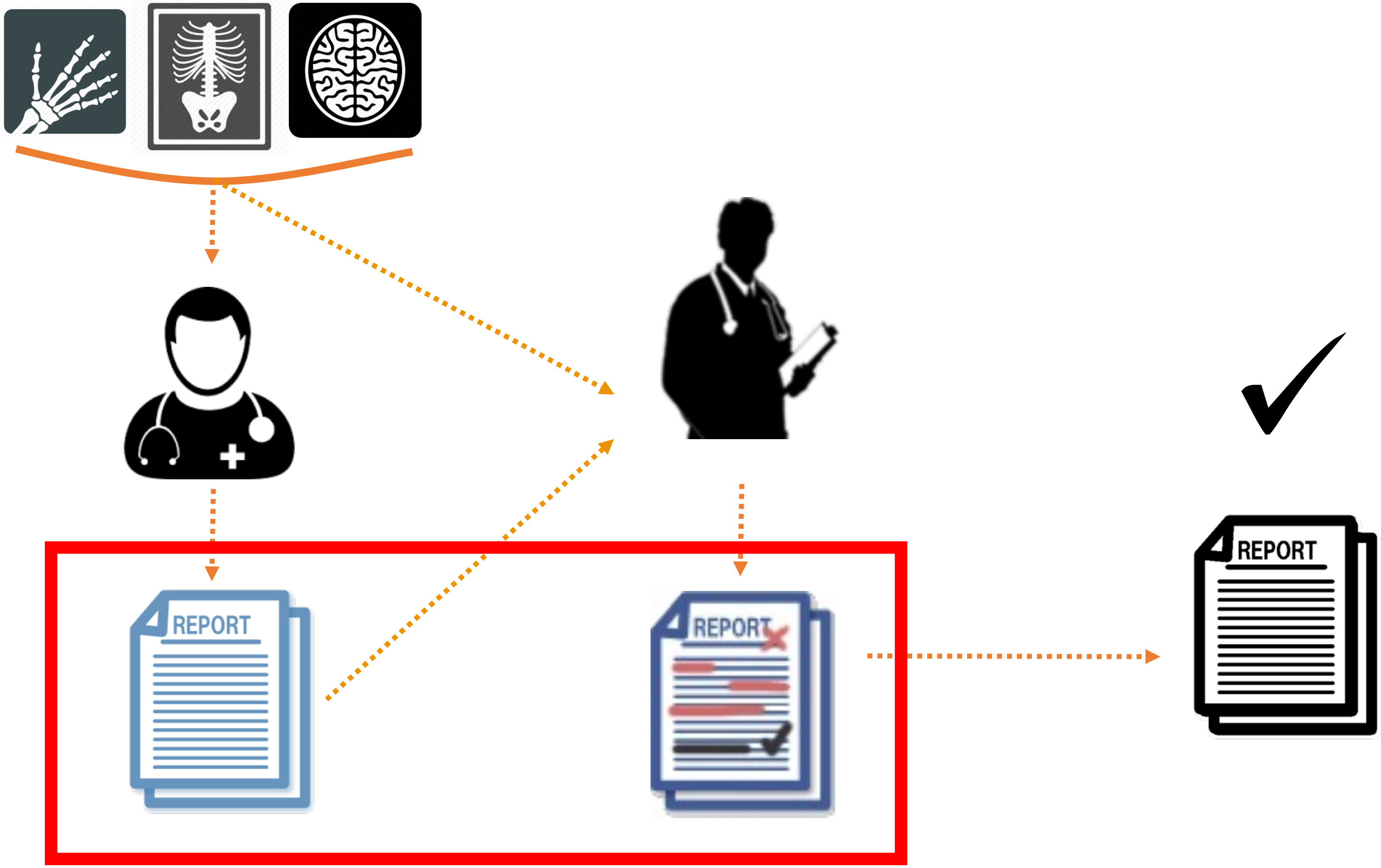
MedStar Health Research Institute

MedStar Georgetown University Hospital

Outline

- Motivation
- Framework
- Evaluation
- Conclusions

Overview of Clinical Workflow

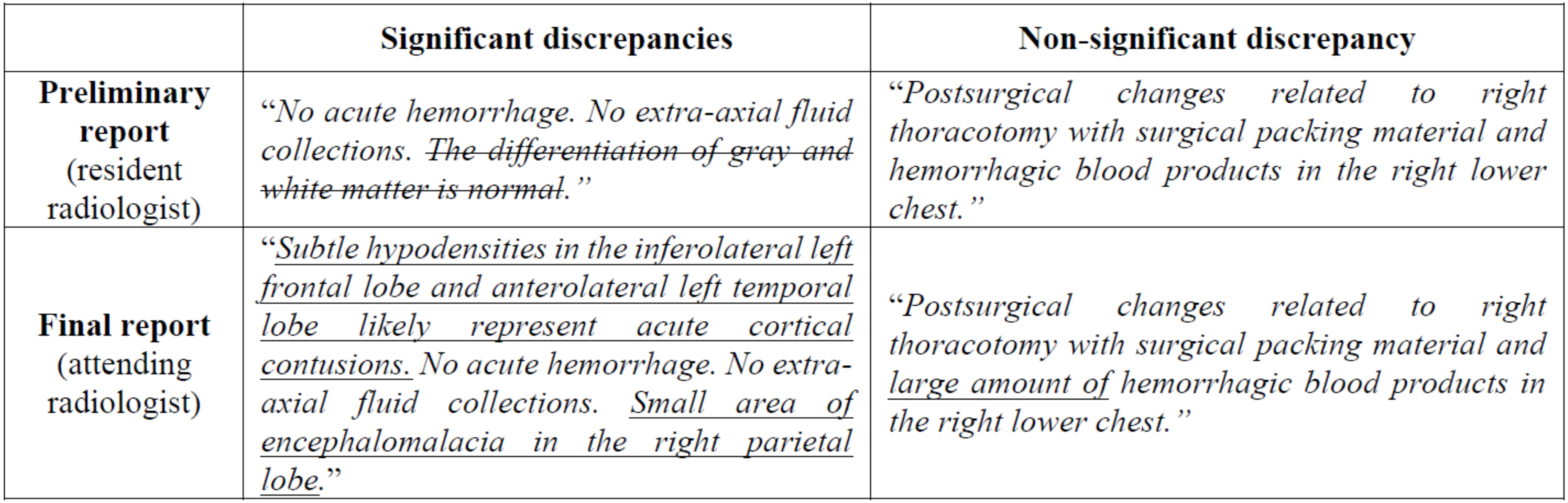


Example

Initial Report	Final Report
No acute hemorrhage. No extra-axial fluid collections. ~~The differentiation of gray and white matter is normal~~.	"Subtle hypodensities in the inferolateral left frontal lobe and anterolateral left temporal lobe likely represent acute cortical contusions. No acute hemorrhage. No extra-axial fluid collections. Small area of encephalomalacia in the right parietal lobe."

Introduction

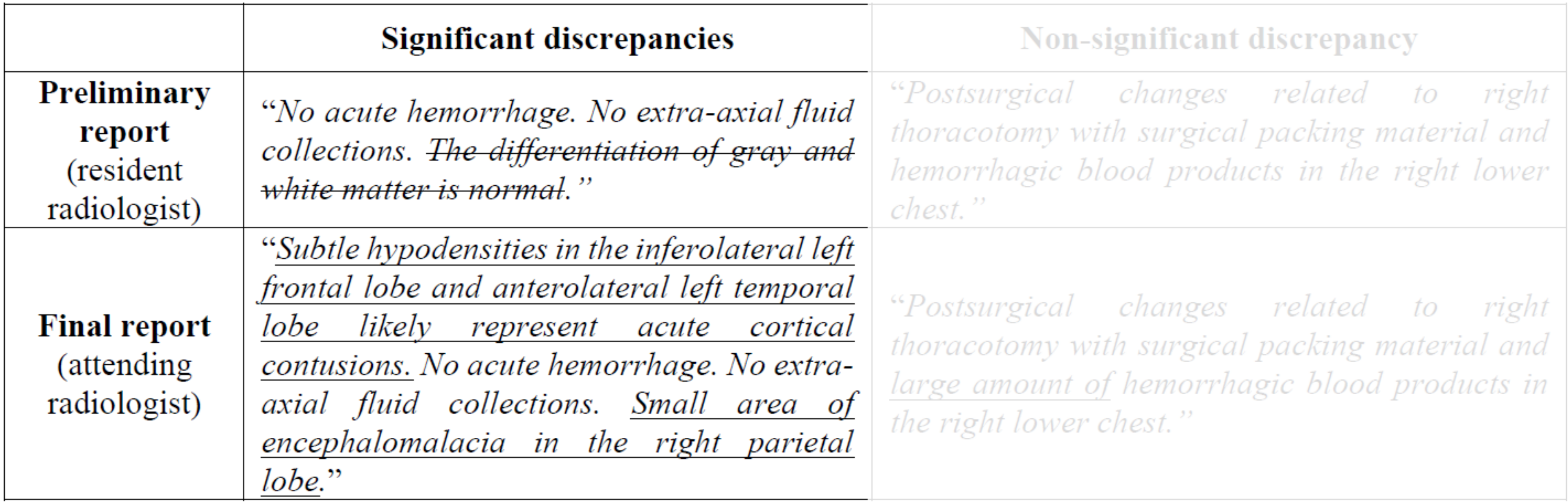
- Two types of discrepancies



	Significant discrepancies	Non-significant discrepancy
Preliminary report (resident radiologist)	*"No acute hemorrhage. No extra-axial fluid collections. ~~The differentiation of gray and white matter is normal~~."*	*"Postsurgical changes related to right thoracotomy with surgical packing material and hemorrhagic blood products in the right lower chest."*
Final report (attending radiologist)	*"Subtle hypodensities in the inferolateral left frontal lobe and anterolateral left temporal lobe likely represent acute cortical contusions. No acute hemorrhage. No extra-axial fluid collections. Small area of encephalomalacia in the right parietal lobe."*	*"Postsurgical changes related to right thoracotomy with surgical packing material and large amount of hemorrhagic blood products in the right lower chest."*

Introduction

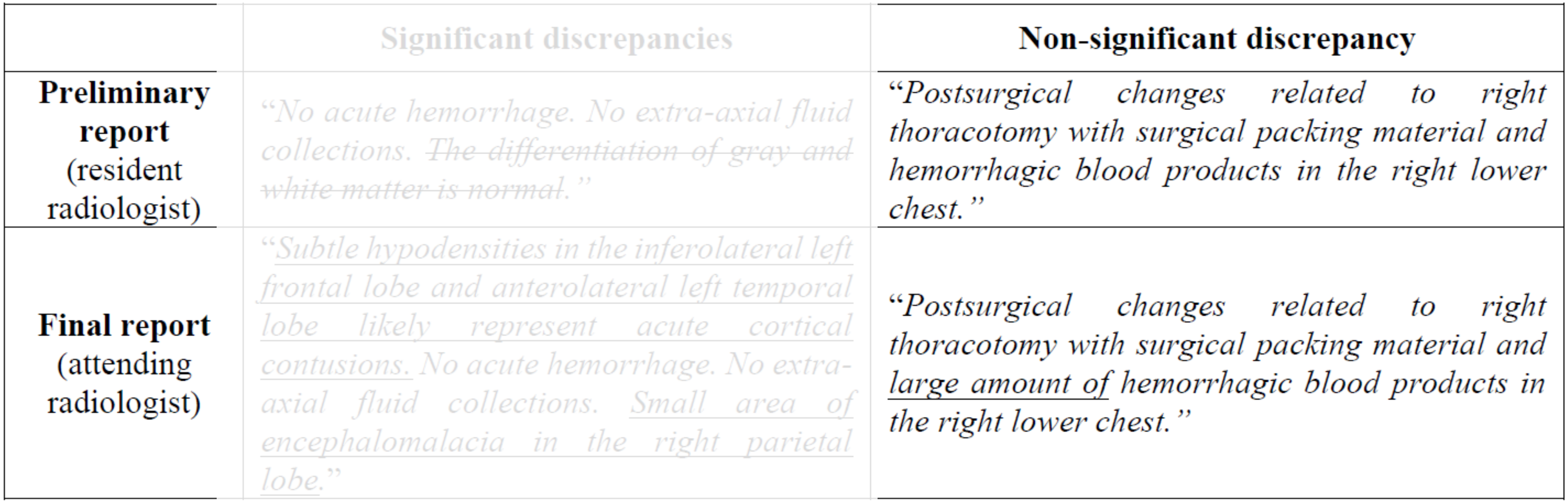
- Two types of discrepancies



	Significant discrepancies	Non-significant discrepancy
Preliminary report (resident radiologist)	*"No acute hemorrhage. No extra-axial fluid collections. ~~The differentiation of gray and white matter is normal~~."*	*"Postsurgical changes related to right thoracotomy with surgical packing material and hemorrhagic blood products in the right lower chest."*
Final report (attending radiologist)	*"Subtle hypodensities in the inferolateral left frontal lobe and anterolateral left temporal lobe likely represent acute cortical contusions. No acute hemorrhage. No extra-axial fluid collections. Small area of encephalomalacia in the right parietal lobe."*	*"Postsurgical changes related to right thoracotomy with surgical packing material and large amount of hemorrhagic blood products in the right lower chest."*

Introduction

- Two types of discrepancies



	Significant discrepancies	Non-significant discrepancy
Preliminary report (resident radiologist)	*"No acute hemorrhage. No extra-axial fluid collections. ~~The differentiation of gray and white matter is normal.~~"*	*"Postsurgical changes related to right thoracotomy with surgical packing material and hemorrhagic blood products in the right lower chest."*
Final report (attending radiologist)	*"Subtle hypodensities in the inferolateral left frontal lobe and anterolateral left temporal lobe likely represent acute cortical contusions. No acute hemorrhage. No extra-axial fluid collections. Small area of encephalomalacia in the right parietal lobe."*	*"Postsurgical changes related to right thoracotomy with surgical packing material and large amount of hemorrhagic blood products in the right lower chest."*

Problem

- Significant discrepancies are important
  - In the patient care and resident's education
- Manual surveillance is difficult
- Previous work: Using wording differences (Kalaria, et al. 2015)
  - Not-accurate: Many wordings are due to style changes and do not reflect misinterpretations or misdiagnoses

This work

We propose a framework for accurate identification of significant reports

Outline

- Motivation ✓
- Framework ←
- Evaluation
- Conclusions

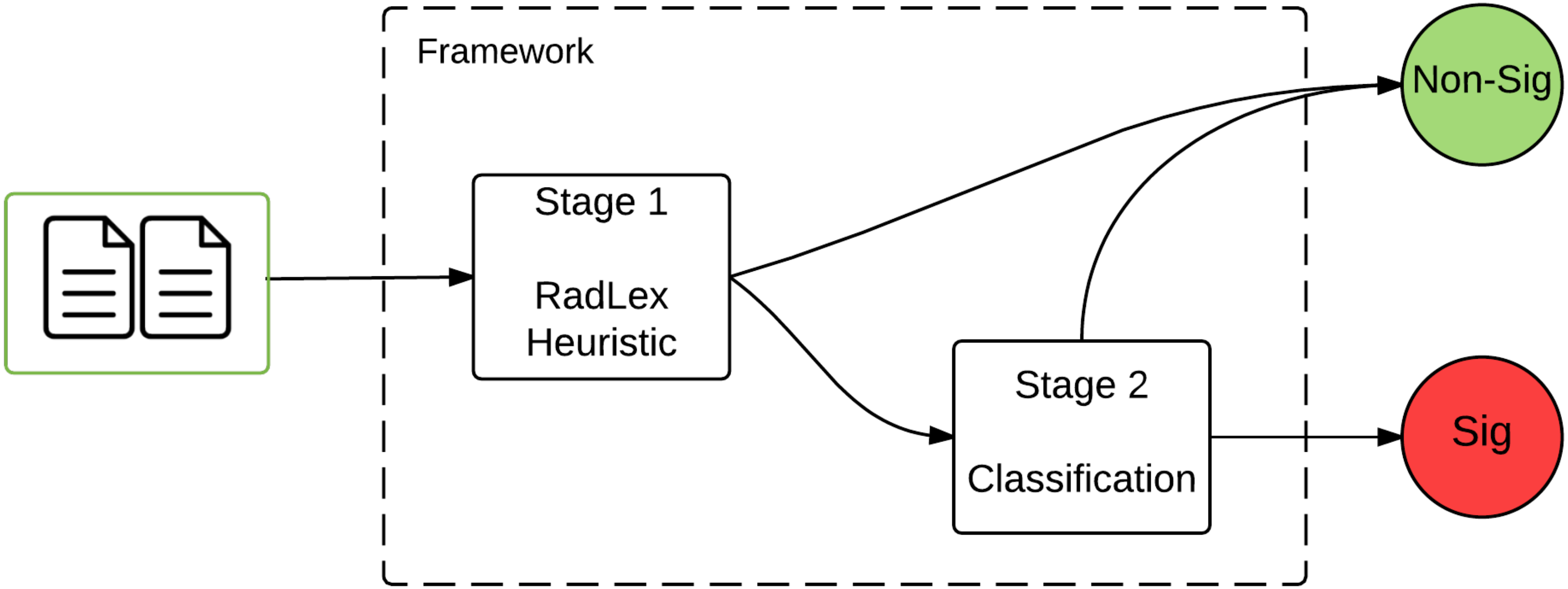
MedStar Institute for Innovation

1789 GEORGETOWN UNIVERSITY

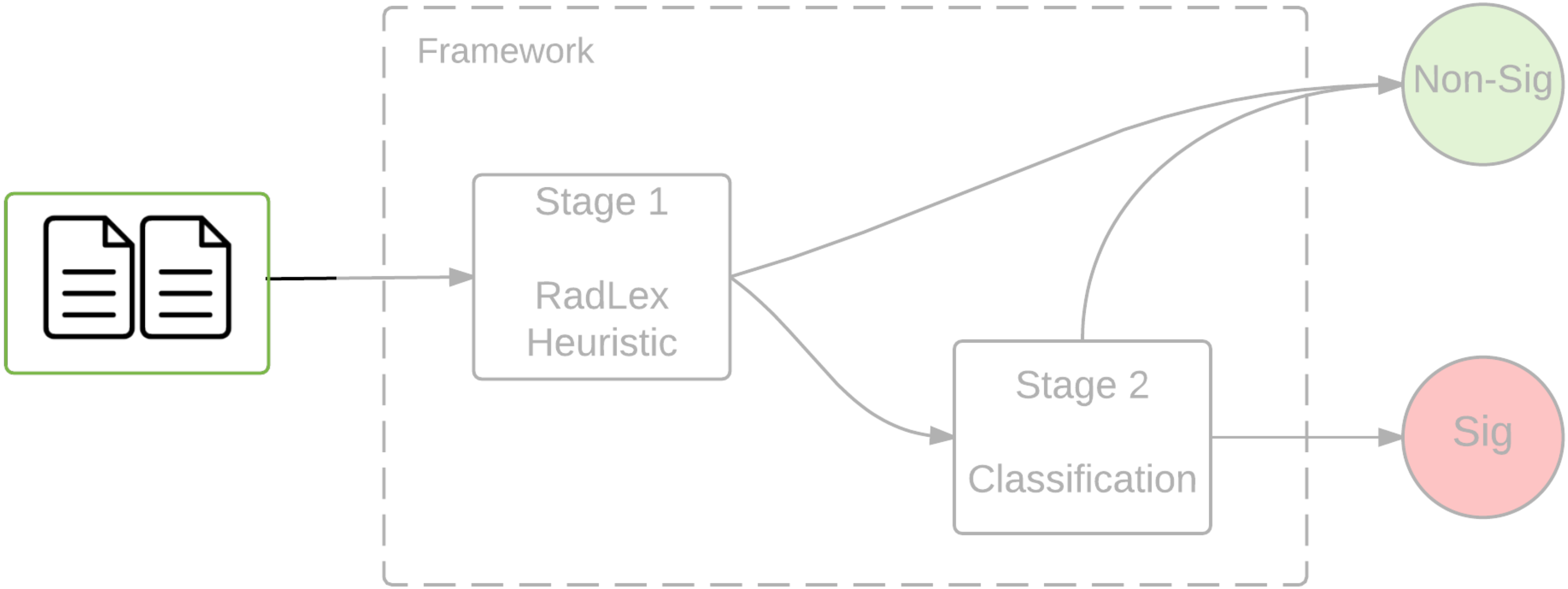
Data

- Collection of annotated radiology reports with discrepancies obtained from a large urban hospital for evaluation.
  - Set of 350 reports
- Two sections for each report:
  - Findings: contains the full interpretation of the radiology examination
  - Impression: highlights important aspects of the report.

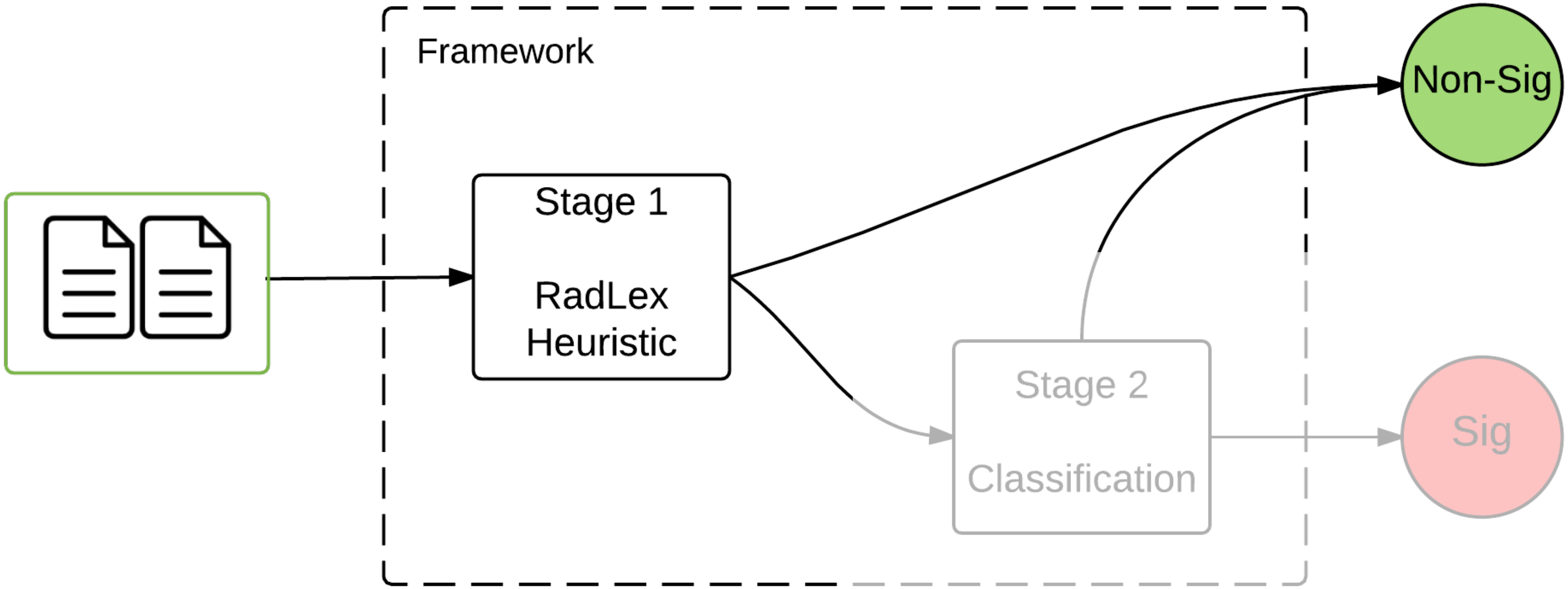
Overview


Framework
Stage 1
RadLex Heuristic
Stage 2
Classification
Non-Sig
Sig

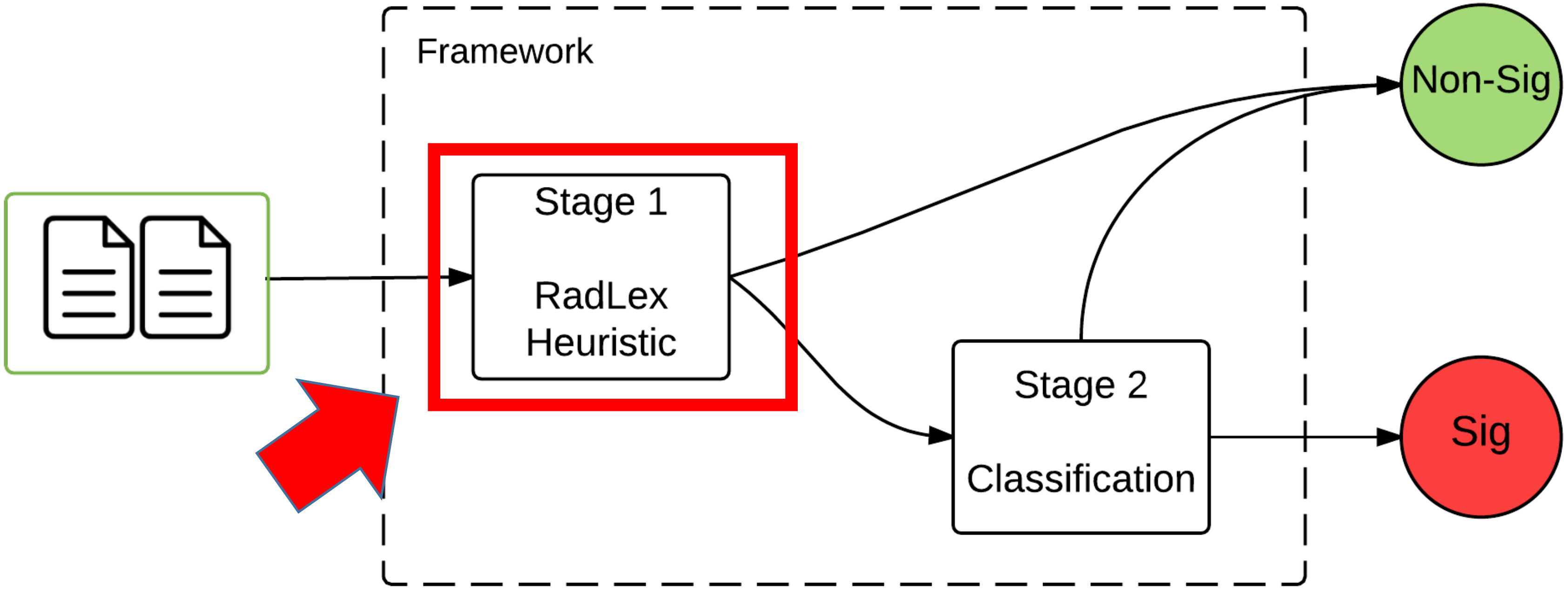
Overview


Framework
Stage 1
RadLex
Heuristic
Stage 2
Classification
Non-Sig
Sig

Overview


Framework
Stage 1
RadLex
Heuristic
Stage 2
Classification
Non-Sig
Sig

Overview


Framework
Stage 1
RadLex
Heuristic
Stage 2
Classification
Non-Sig
Sig

Stage 1: Radlex heuristic

- If the Radlex terms are identical and negations are consistent, classify as non-significant
- Compare domain specific concepts (Radlex Ontology)

Prem. : "… There is a diffuse, dense, airspace opacity occupying most of the …"
Final: : "… diffuse, dense, airspace opacity occupying left frontal …"

- Negations:

Prem. : "… hypodensities in the inferolateral left frontal lobe …"
Final: "… **no** hypodensity in the inferolateral left frontal lobe …"

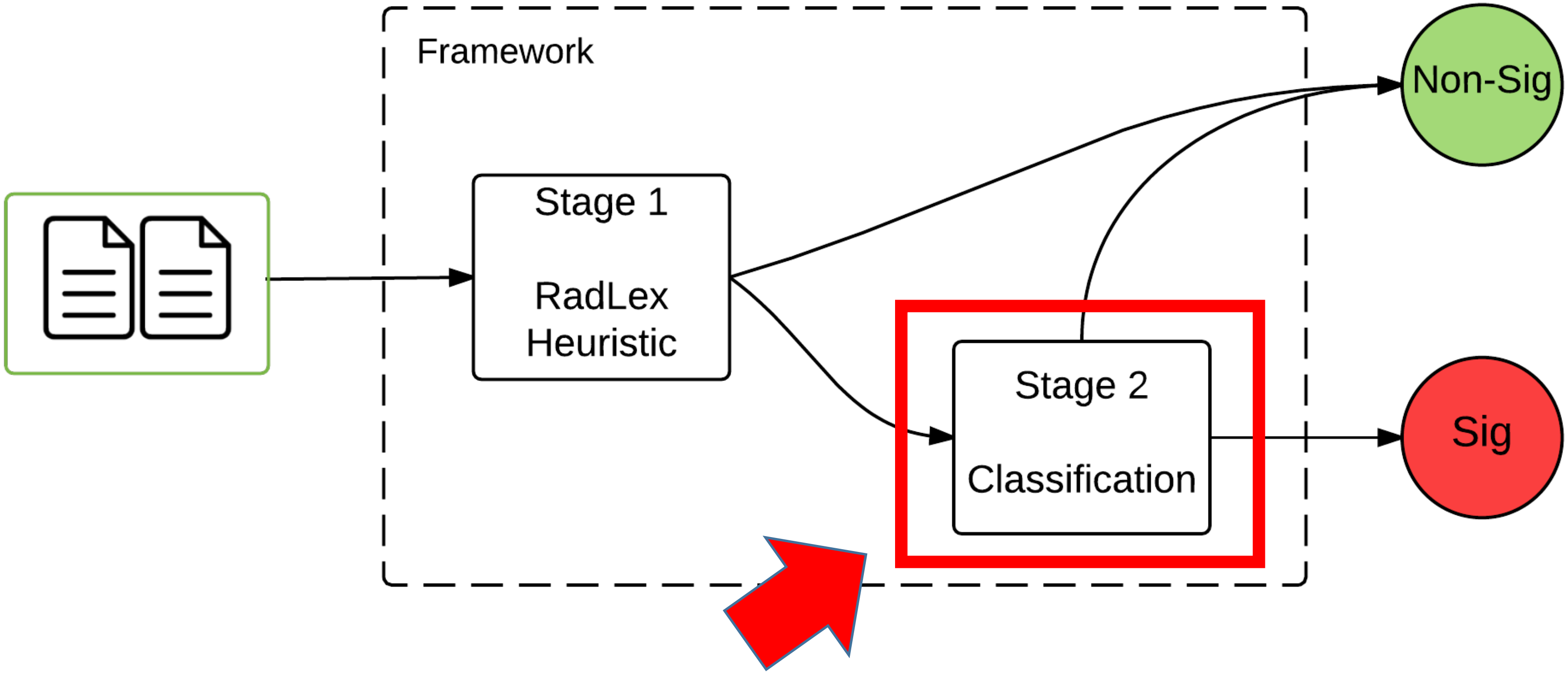
Stage-1: Radlex heuristic

		Radlex	Human A	Human B
Non-Significant	Radlex	1.0	0.964	0.942
	Human A	0.946	1.0	0.906
	Human B	0.942	0.906	1.0
Count=139	Fleiss $\kappa = 0.880$			

Stage-1: Radlex heuristic

		Radlex	Human A	Human B
Non-Significant	**Radlex**	1.0	0.964	0.942
	Human A	0.946	1.0	0.906
	Human B	0.942	0.906	1.0
Count=139	Fleiss $\kappa = 0.880$			
Significant	**Radlex**	1.0	0.557	0.492
	Human A	0.557	1.0	0.934
	Human B	0.492	0.934	1.0
Count=61	Fleiss $\kappa = 0.468$			

Overview


Framework
Stage 1
RadLex Heuristic
Stage 2
Classification
Non-Sig
Sig

Stage 2: Classification

- Use textual features designed for capturing the differences between the reports.
- Feed this features to a classifier

Features

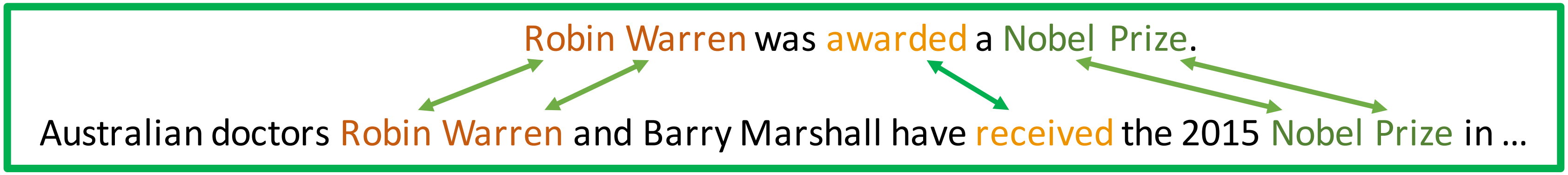
- Surface Textual Features
  - Character, word and sentence differences
- Summarization evaluation features
  - ROUGE: Evaluation metric based on text overlaps
    - Take the final report as the gold standard and compute the ROUGE score of the preliminary report
    - Higher scores → Differences are less significant → Higher quality of preliminary report

Features

- ROUGE-N:
  - N-Gram precision and recall
- ROUGE-L:
  - Sequence differences (LCS)
- ROUGE-S:
  - Skip-Bigram co-occurrence

Features

- Machine Translation Evaluation Metrics
  - Take the final report as gold-standard
  - Evaluate the quality of the preliminary report
    - BLEU: Similar to Rouge-N, except being precision-oriented. With brevity penalty
    - Word Error Rate (WER)
    - METEOR: Based on alignment, considers synonyms



Robin Warren was awarded a Nobel Prize.

Australian doctors Robin Warren and Barry Marshall have received the 2015 Nobel Prize in …

Features

- Readability assessment
  - Quantify and compare the reporting stylistic characteristic of the reports
    - Automated Readability Index (ARI)
    - Simple Measure of Gobbledygook (SMOG) index
    - Average of phrase counts

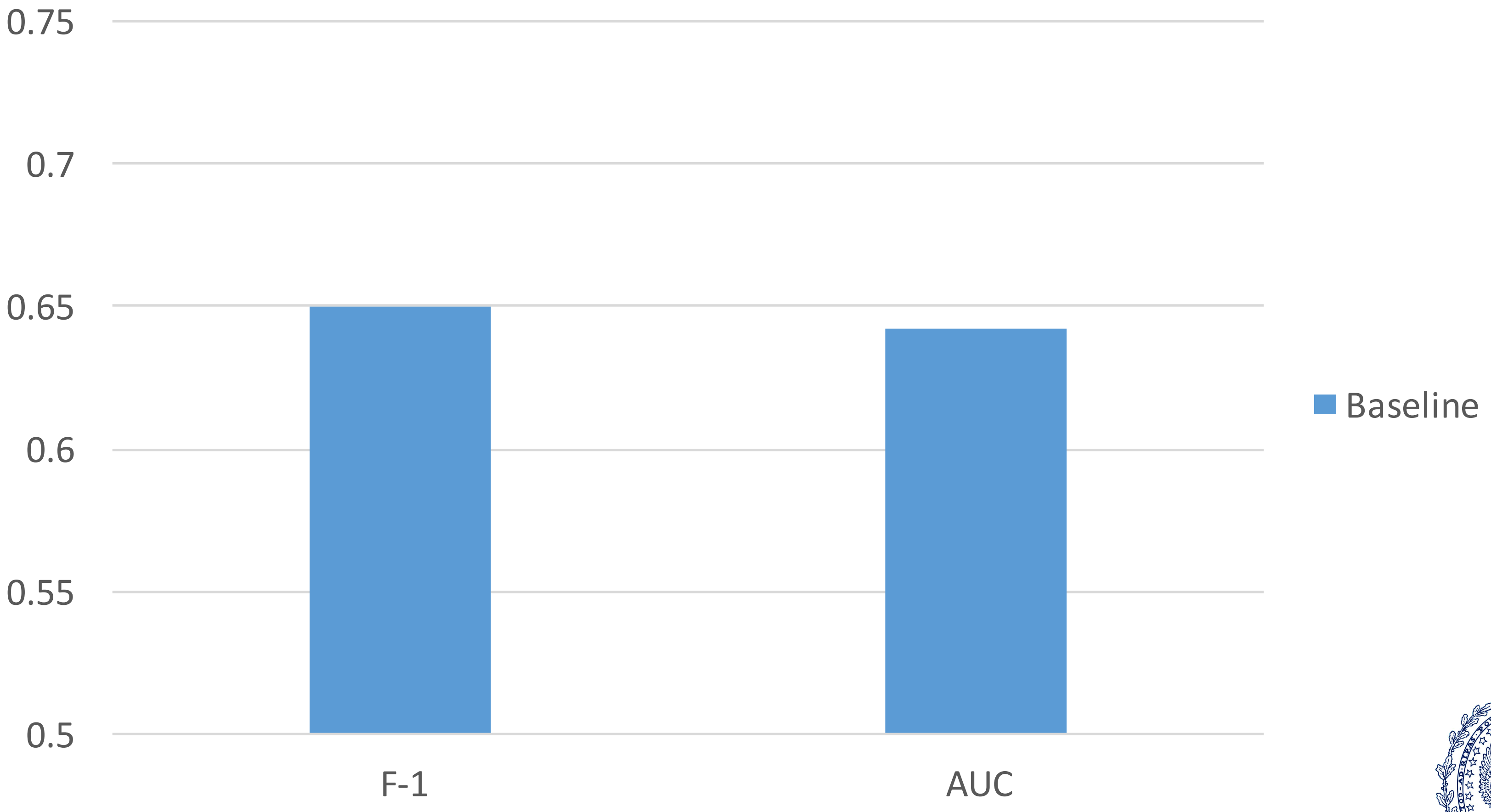
Outline

- Motivation ✓
- Framework ✓
- Evaluation ⬅
- Conclusions

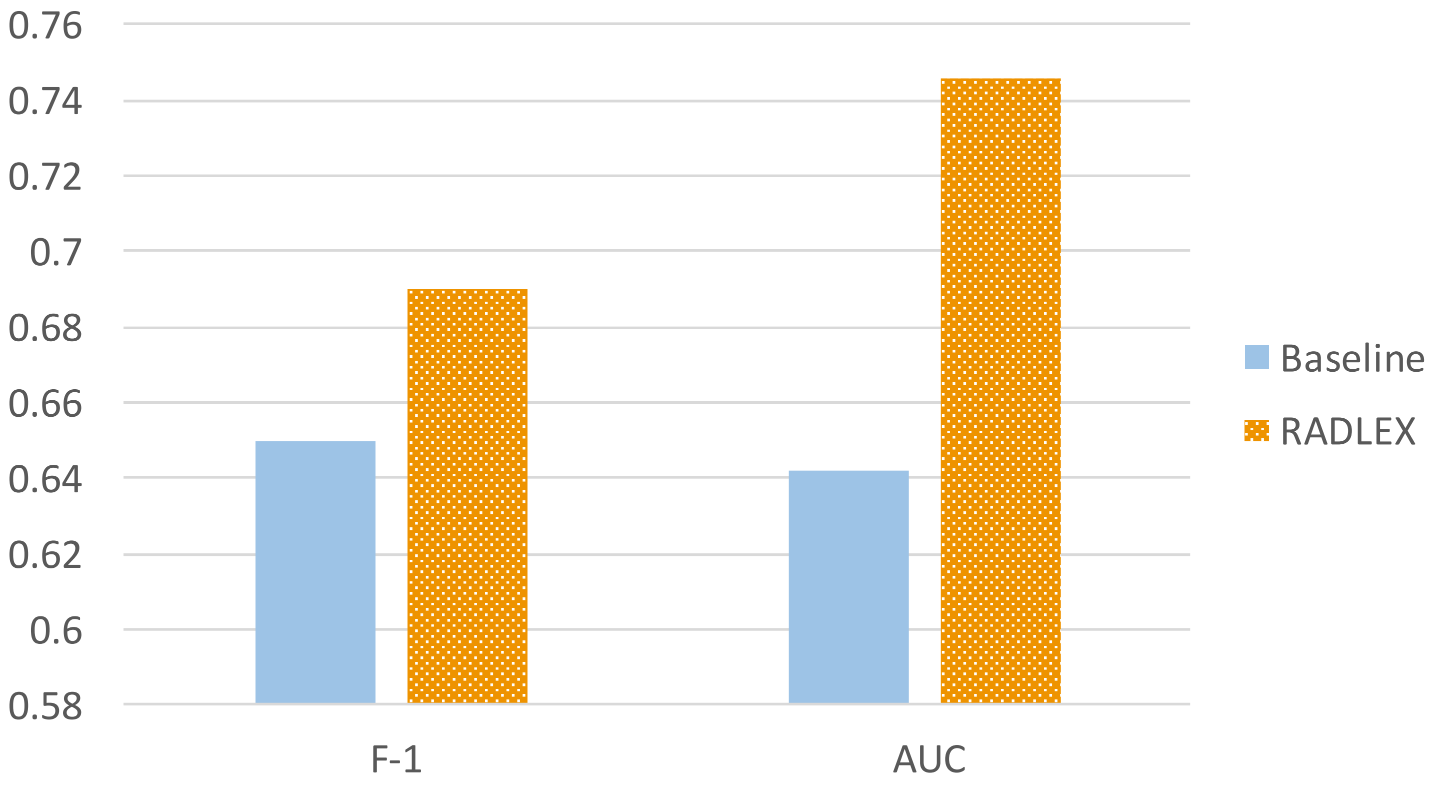
Evaluation

- Stage 1 – Radlex Heuristic
  - 200 Manually annotated reports
- Stage 2 – Classification approach
  - 150 Manually annotated reports

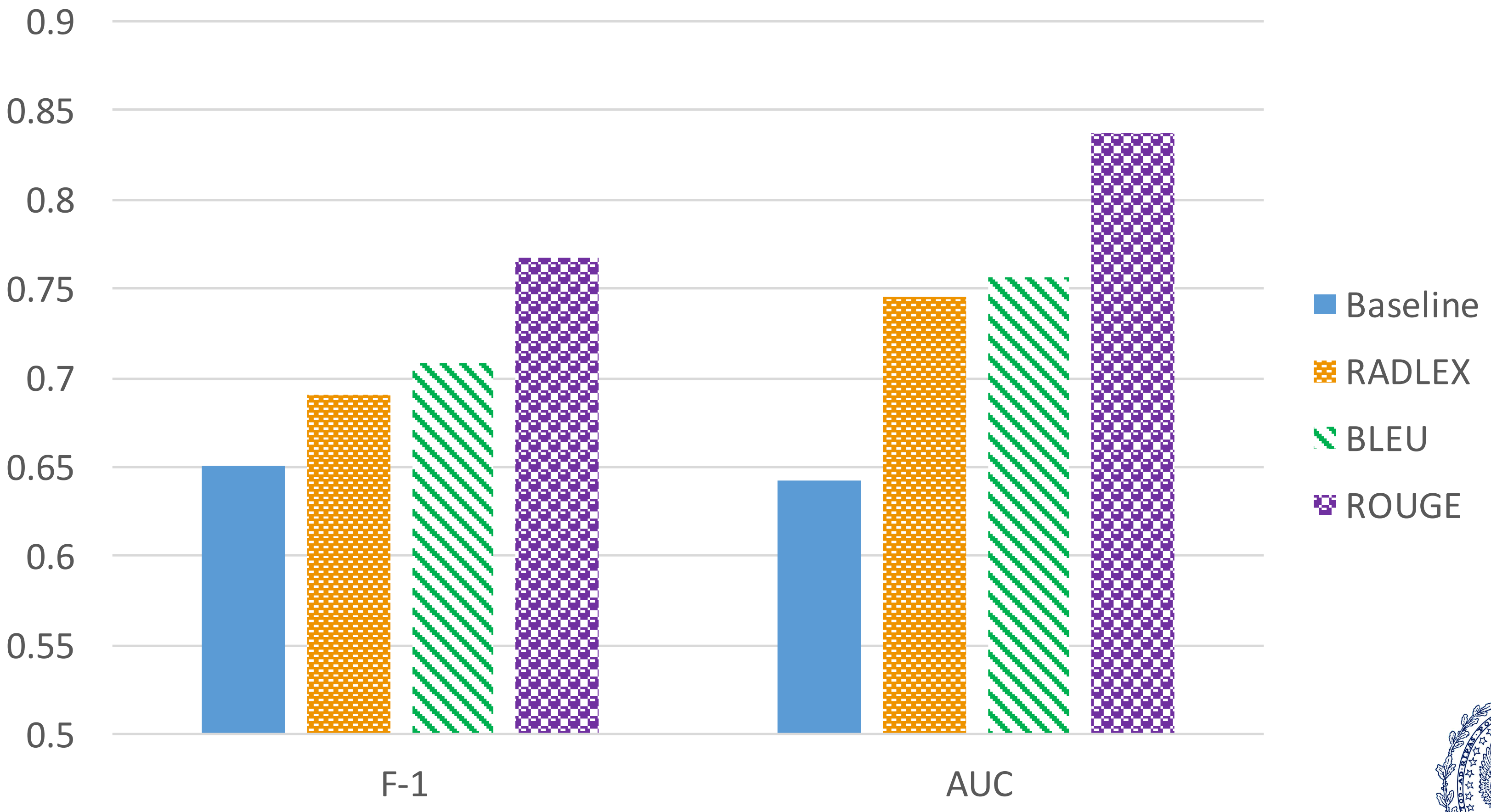
Results



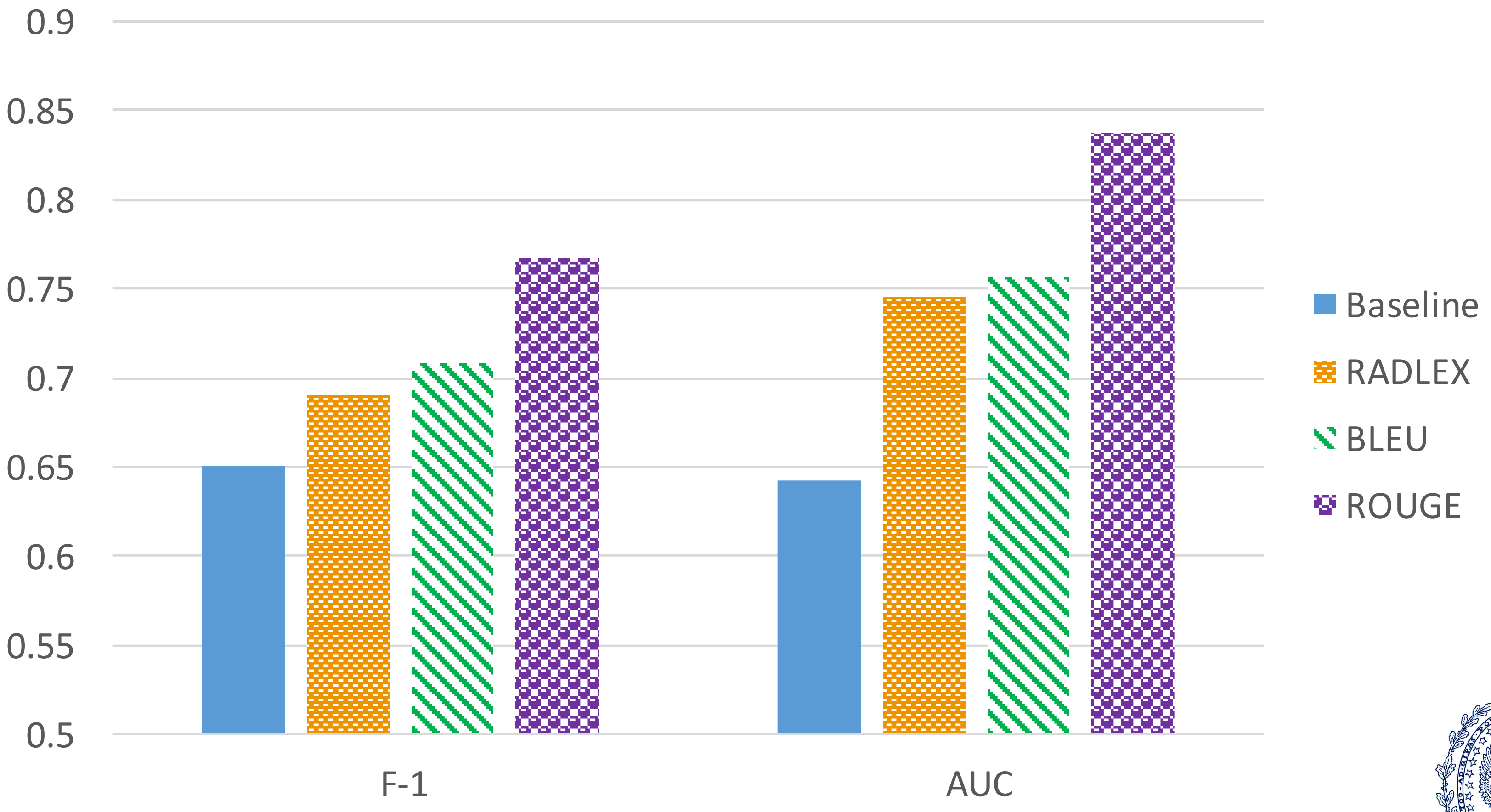
Results: Individual Features



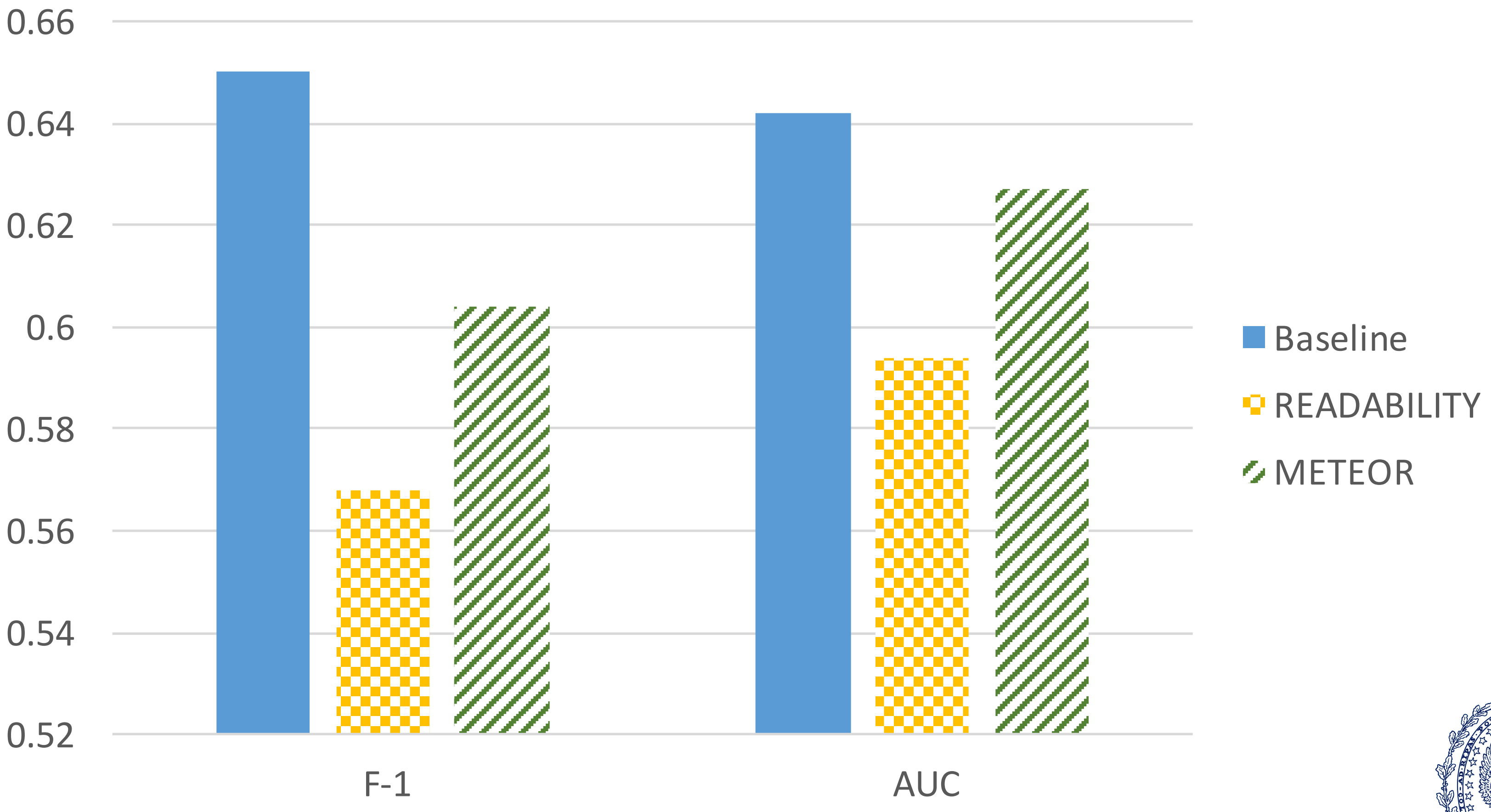
Results: Individual Features



Results: Individual Features

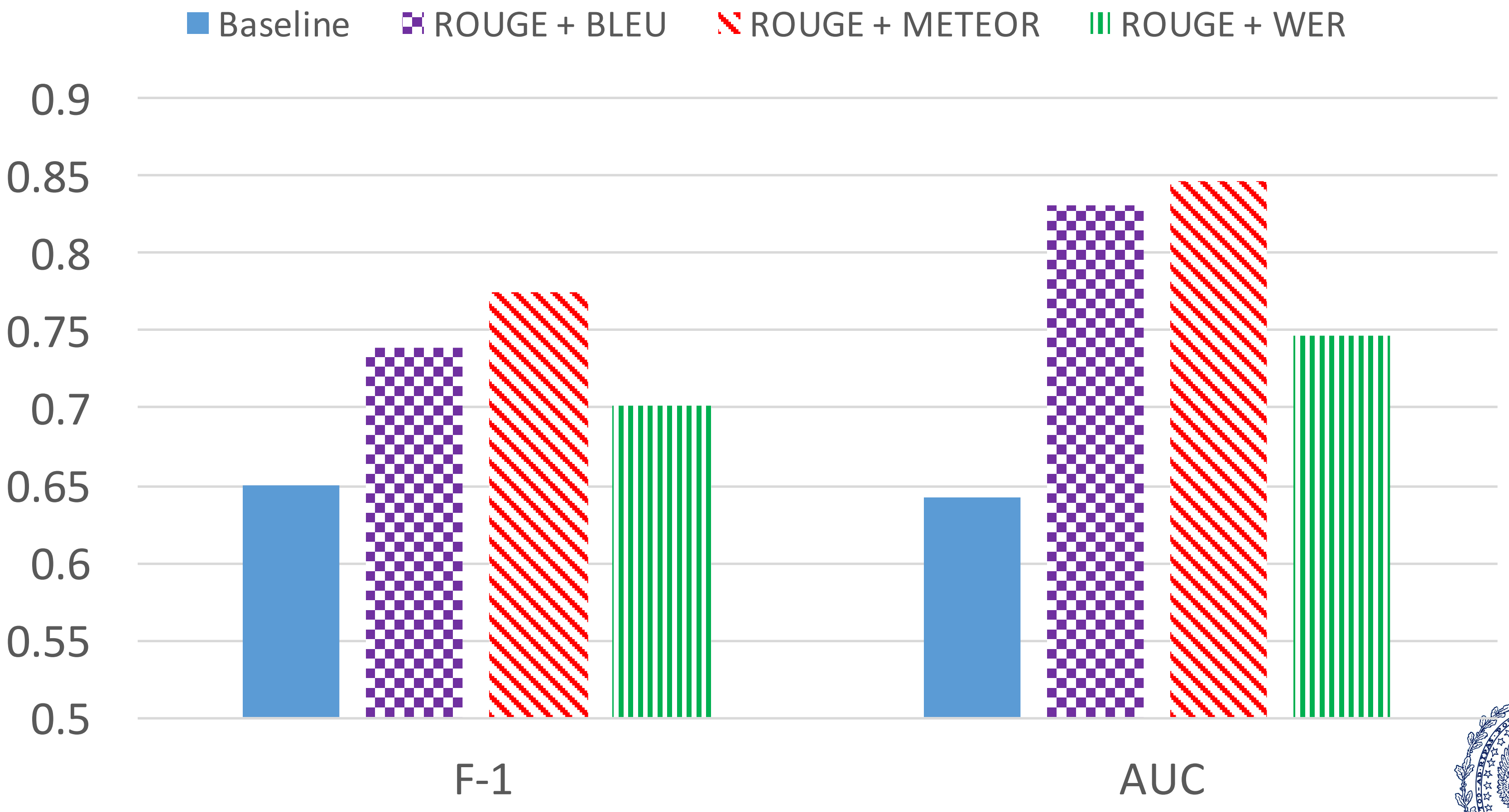


Results: Individual Features

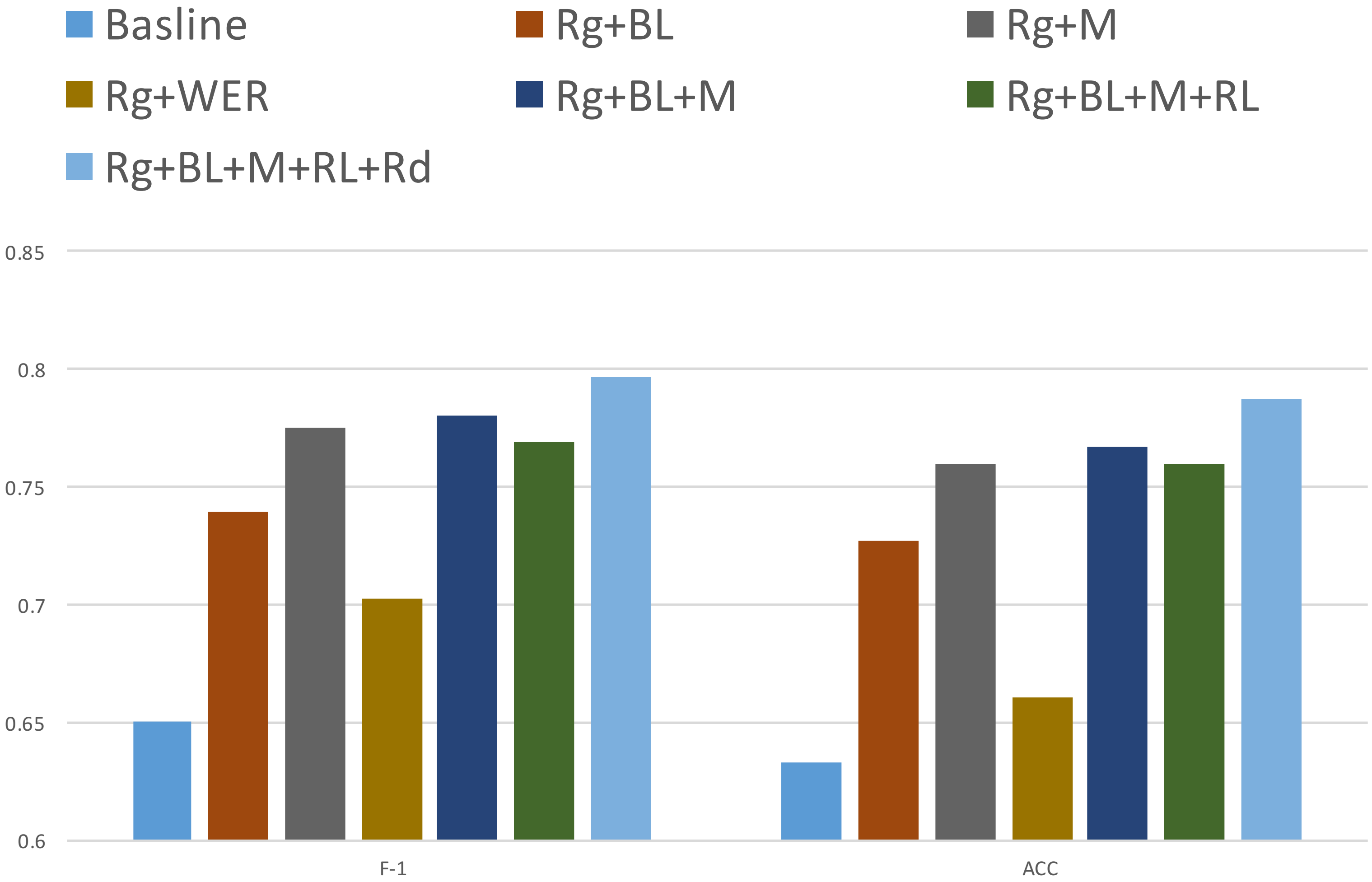


Feature Combinations

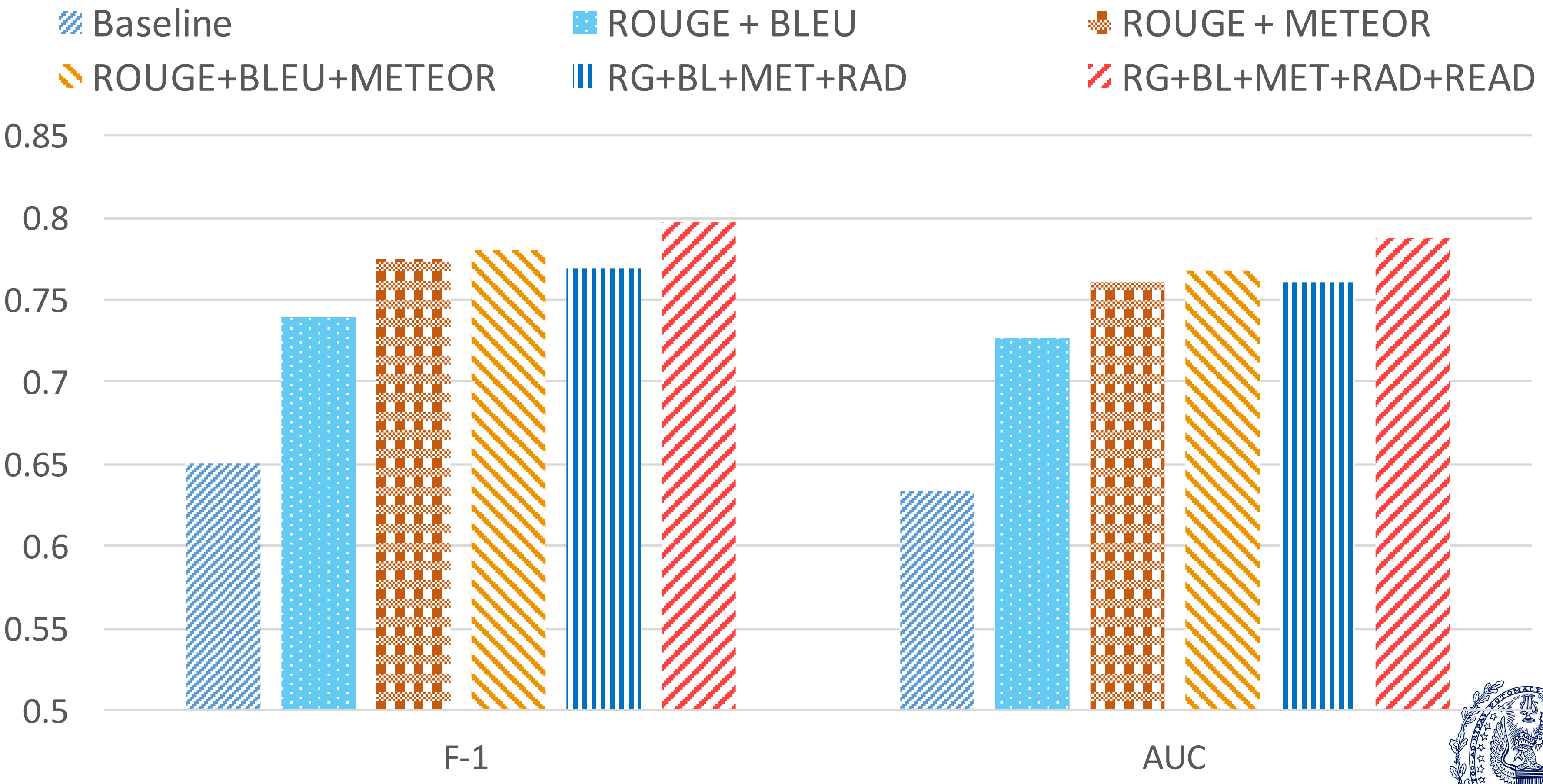
Summarization and MT



Feature Combinations

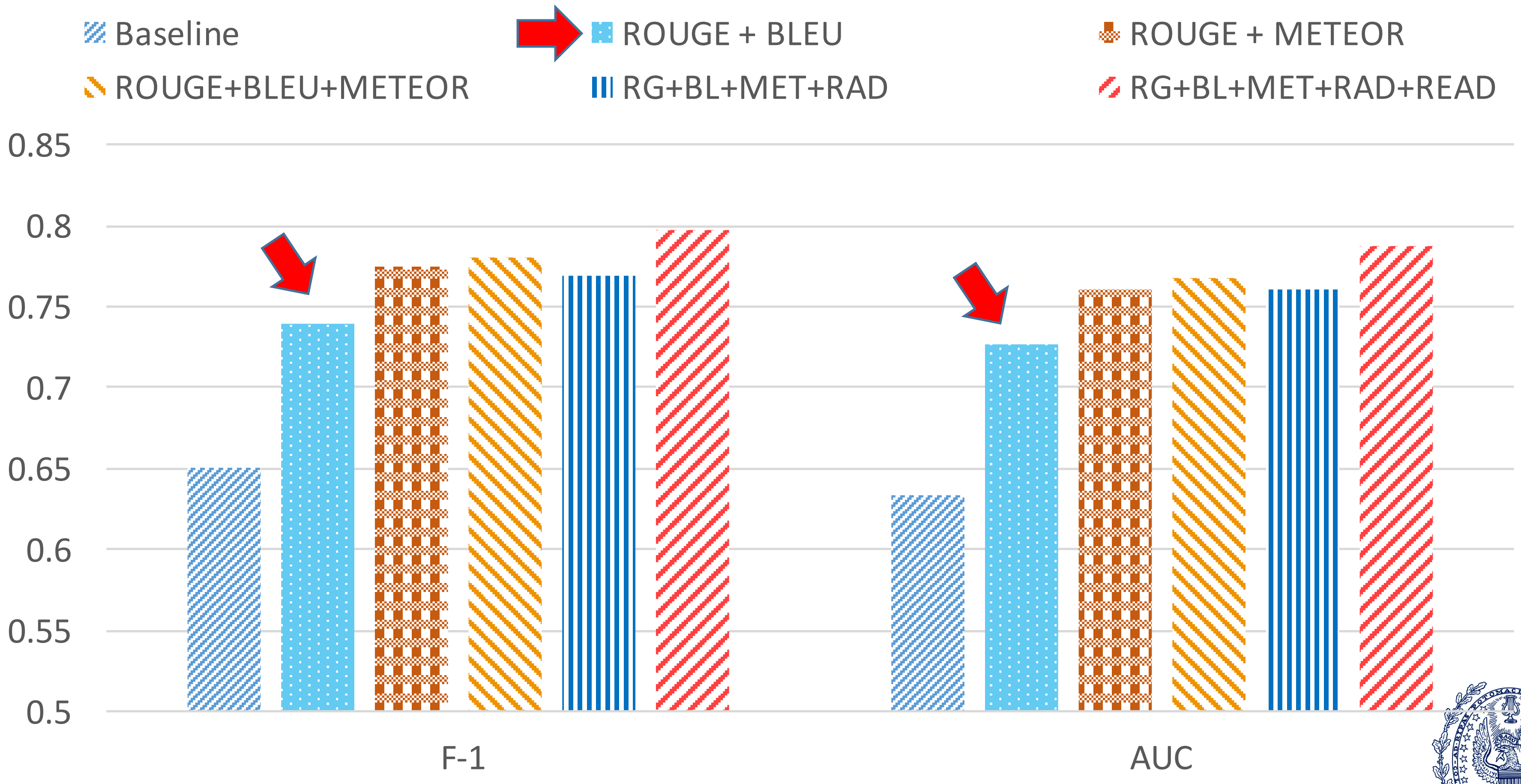


Feature Combinations



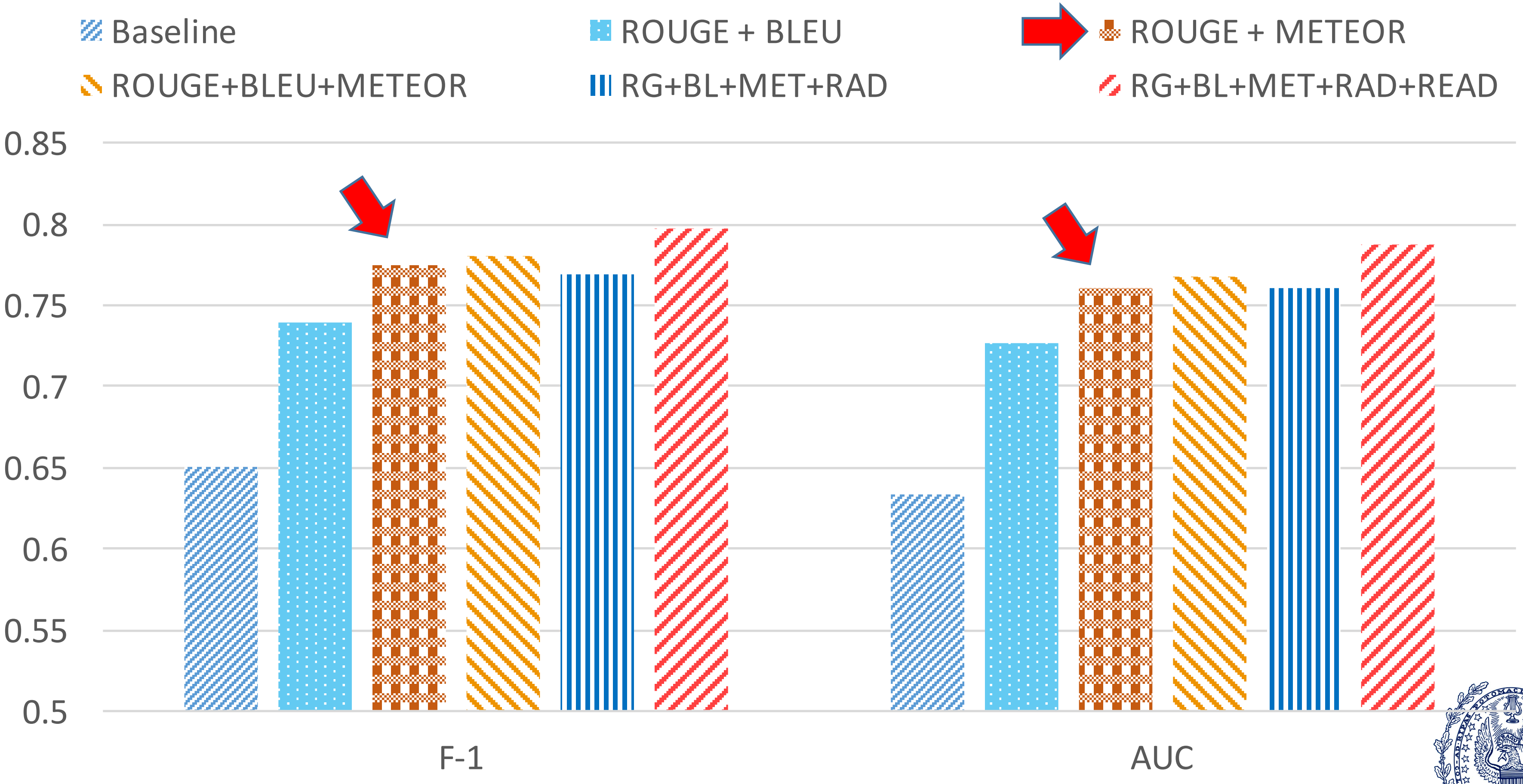
MedStar Institute for Innovation

Feature Combinations

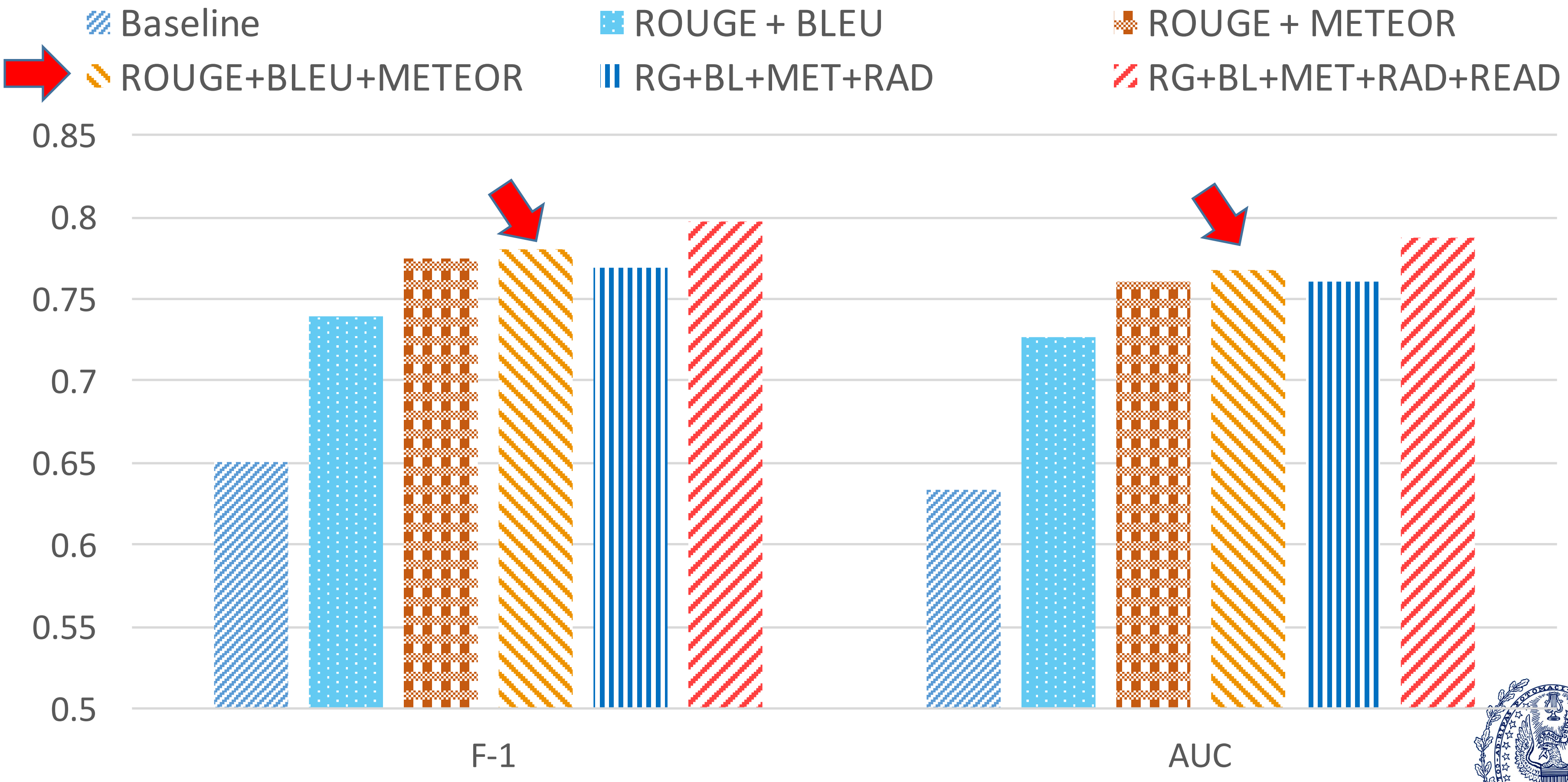


MedStar Institute for Innovation

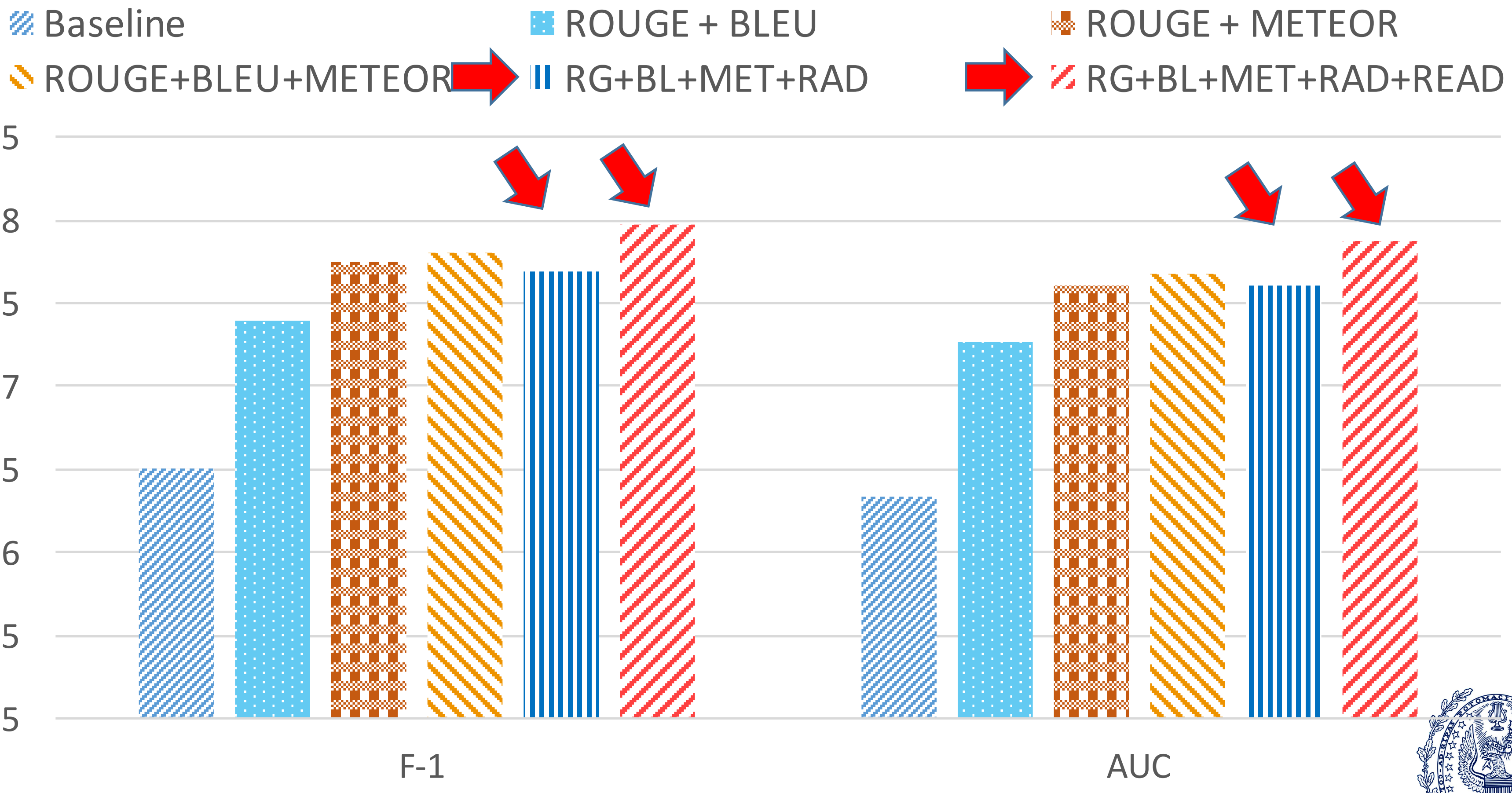
Feature Combinations



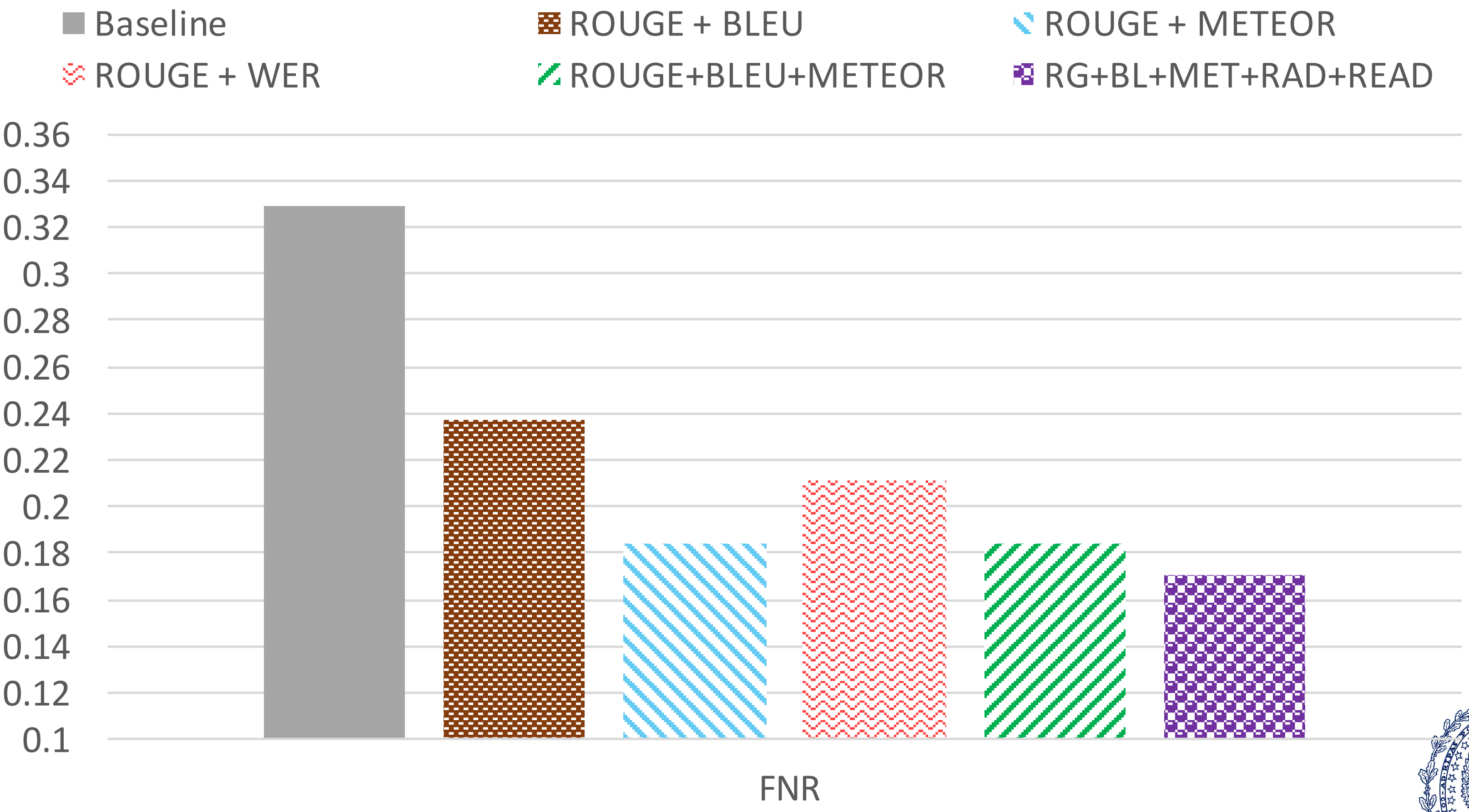
Feature Combinations



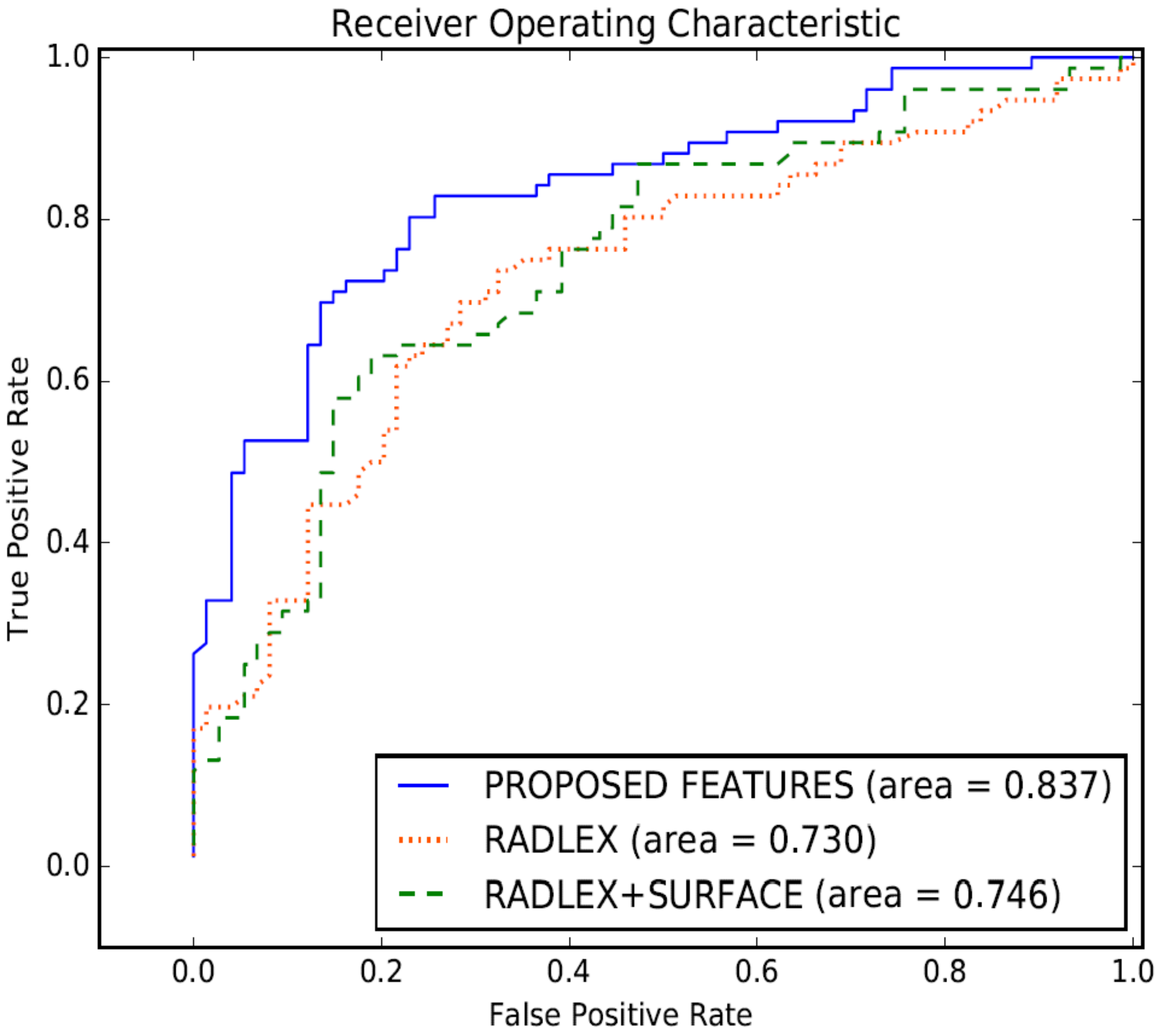
Feature Combinations



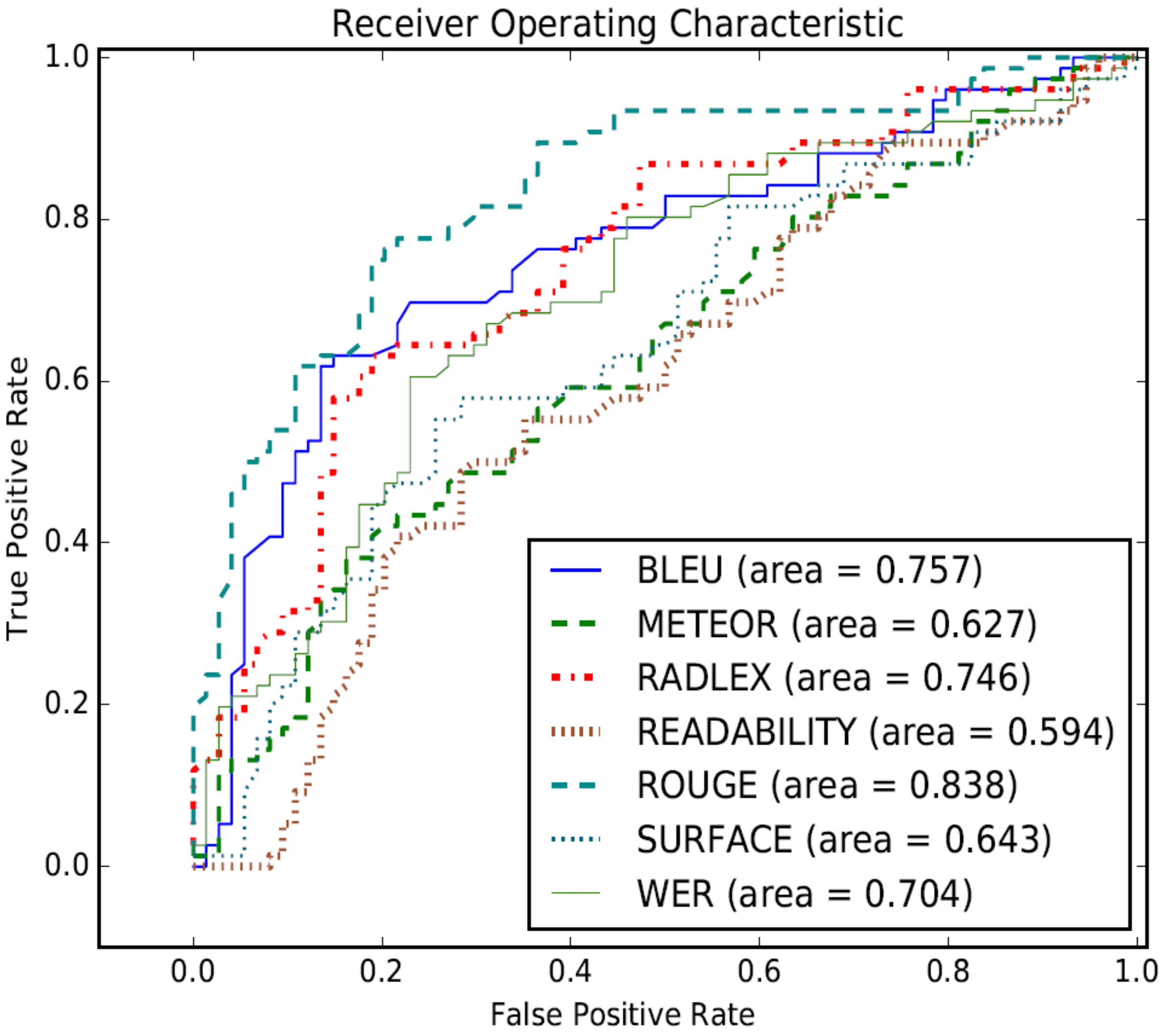
False Negative Rate



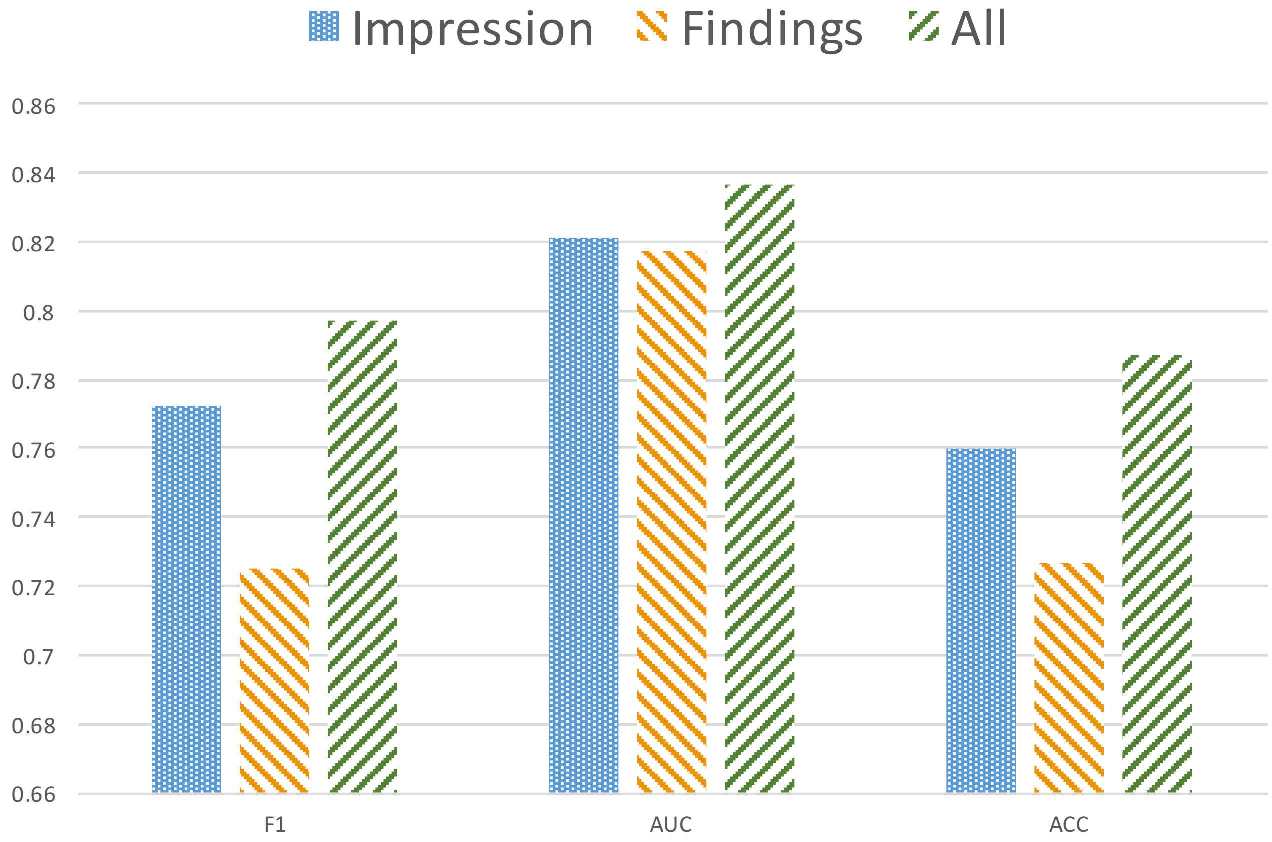
Proposed Features vs Baseline



Feature Comparison



Is summary enough?



Identifying Significance of Discrepancies in Radiology Reports (SDM-DMMH 16)

Error analysis

- False positive cases
  - Unnecessary long length of preliminary reports that were removed in the final version
- False negative cases
  - Very slight change that alters the significance

Prem. Report: Worsening airspace disease at the left base represents aspiration.

Final Report: Worsening airspace disease at the left base could represent aspiration.

Outline

- Motivation ✓
- Framework ✓
- Evaluation ✓
- Conclusions ⬅

Summary

- We can utilize metrics used for evaluation tasks as features for text comparison
- Our two-stage approach effectively identifies the significant discrepancies (79.7 F-1, 17.1 FNR)
- There are special cases that the current features are not designed to handle

Thank you!

Questions?

www.ArmanCohan.com