Ranking Significant Discrepancies in Clinical Reports

Sean MacAvaney, Arman Cohan,
Nazli Goharian, Ross Filice

To appear at ECIR 2020 (short paper)
https://arxiv.org/abs/2001.06674

MedStar Health

Radiology reports contain written components.

Radiological Image(s)

Radiological Note

Background

INDICATION : Peripheral edema.
COMPARISON : None.

Findings

The XXXX examination consists of frontal and lateral radiographs of the chest. The cardiomediastinal contours are within normal limits . Pulmonary vascularity is within normal limits . There is a vague right suprahilar density with elevation of the XXXX fissure most XXXX mild subsegmental atelectasis though superimposed infection can not be entirely excluded. The remaining lungs are clear. The visualized osseous structures and upper abdomen are unremarkable.

Impression

Right upper lobe subsegmental atelectasis.
No evidence of heart failure.

Image and note from the Indiana University Chest X-Ray Collection (de-identified)

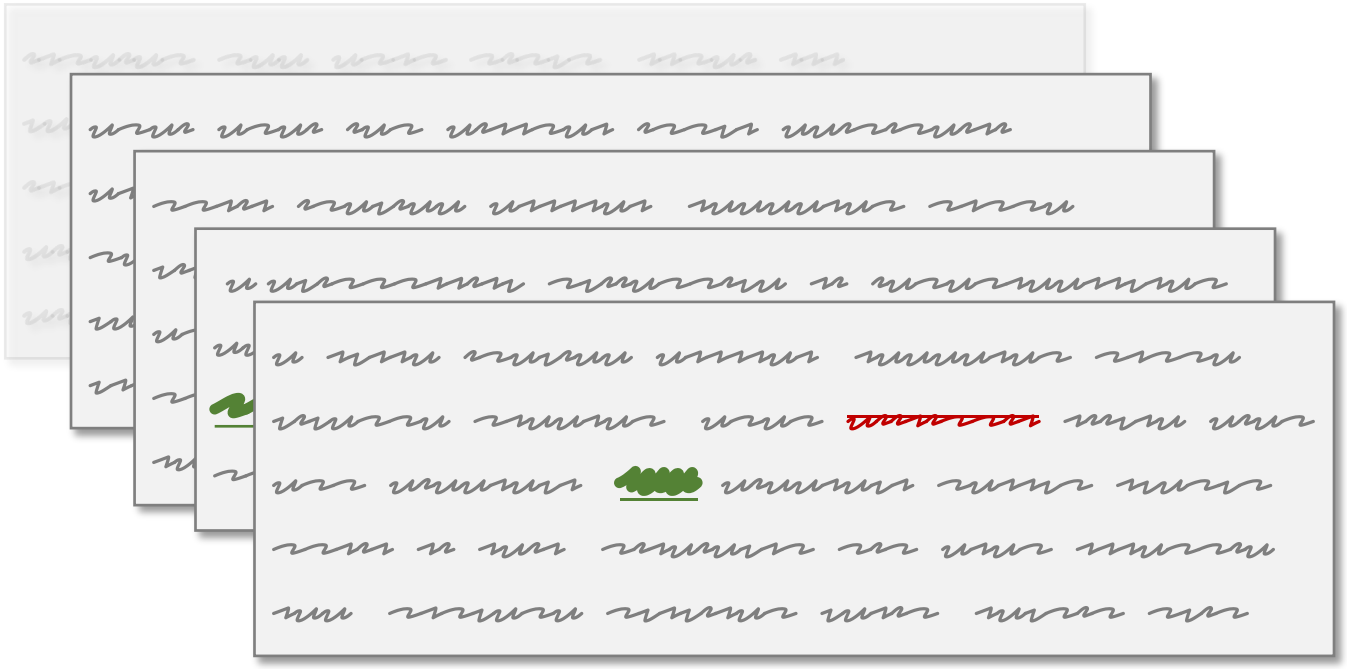
To ensure quality of notes, they often go through revisions.

...with imaging features strongly suggestive of hepatocellular carcinoma (LI-RADS 4) ~~not well discernible~~ **probably present but not conspicuous** on prior examination...

.. 3. Left renal artery: Single with ~~a~~ **slightly early branching** first branch point ~~averaging~~ **averages** 1.9 cm from the left lateral margin of the aorta. Left renal vein: Single without late confluence...

Oftentimes, a medical resident writes the first draft, and then it's edited by an attending radiologist.

Since residents can sometimes write over 100 reports in a week, it can be difficult to look back and find mistakes.

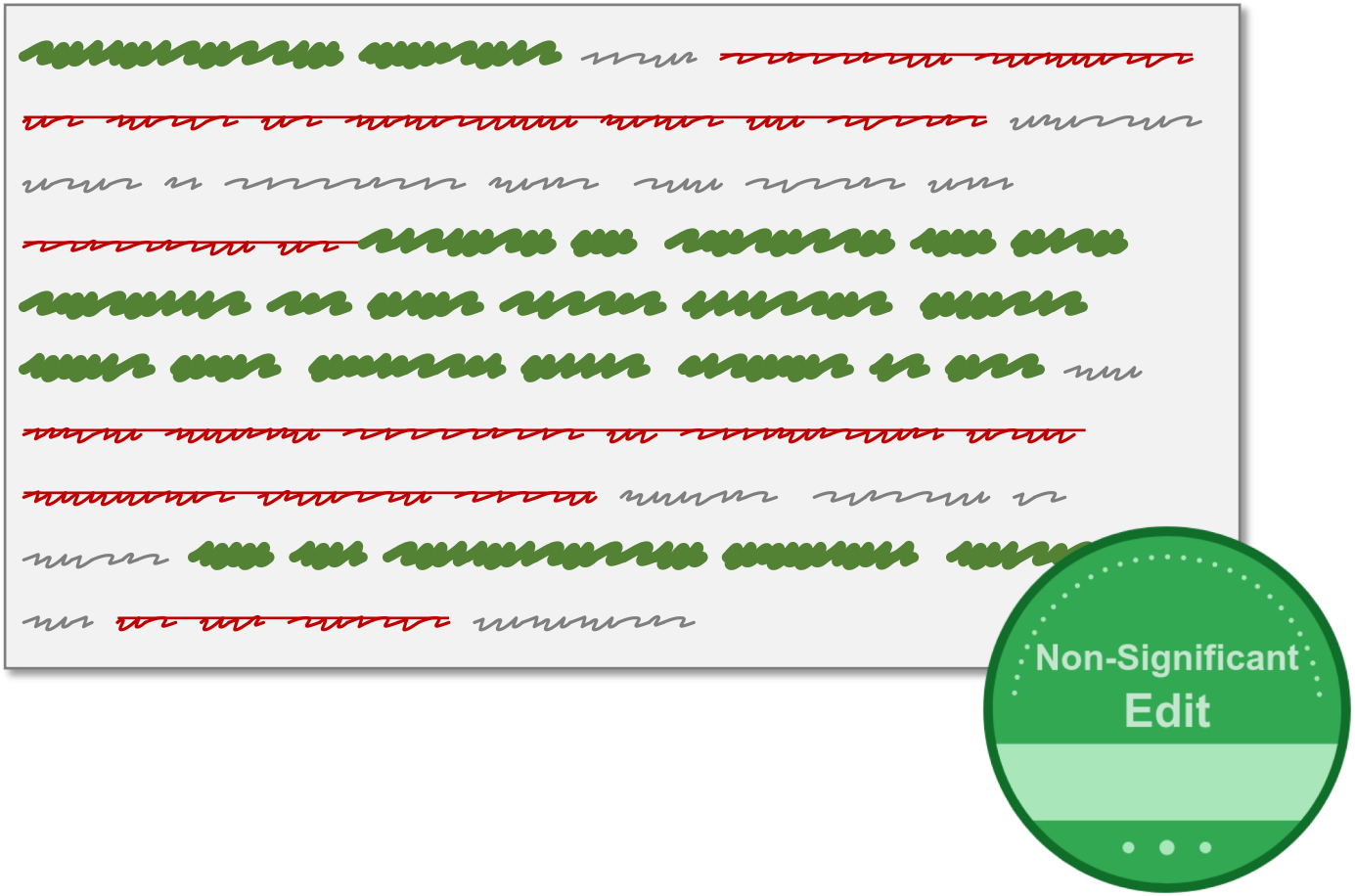


Finding mistakes can reduce risk of future **medical errors**:
a leading cause of harm in hospitals.

We formulate this problem in a ranking setting – rank the reports with the highest degree of discrepancy highest.

This setting allows the resident to find the most important reports first and review as many as they have time for (or until differences become minor).

Challenge: Apparent major changes can be purely stylistic.



Challenge: Apparent Minor changes can have a big impact.

Approach: Model *Importance* and *Similarity* in context

- **Importance:** How much this word (or phrase) would impact the meaning of the report if added/removed?
  - E.g., words like ***not, dislocation, lesion*** may affect meaning significantly, whereas words like ***well, but, evidence*** may not.
- **Similarity:** What's the most similar word (or phrase) within the other version of the text?
  - E.g., maybe not a big deal if ***small*** is replaced by ***mild*** (in the context of ***deformation***).
- We train a model to learn these values for each term in the report based on training data.

Intuition

A **highly-important** term present in the preliminary report that has **low similarity** to any term in the final report may indicate a…

Matching score in preliminary

Importance score in final

$$M_f(p_i)\,I(p_i)$$

Term in preliminary report

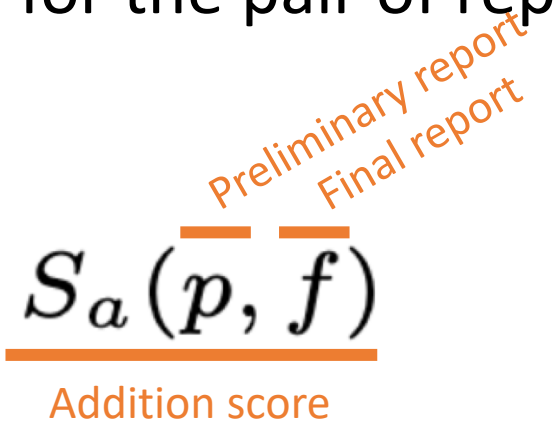
Going from Importance and Similarity to degree of discrepancy.

- Incorporate Importance and Similarity scores into overall “Addition” “Deletion” and “Overlap” scores for the pair of reports.

- We combine these scores using a multi-layer perceptron to produce a final ranking score.

Going from Importance and Similarity to degree of discrepancy.

- Incorporate Importance and Similarity scores into overall “Addition” “Deletion” and “Overlap” scores for the pair of reports.



- We combine these scores using a multi-layer perceptron to produce a final ranking score.

Going from Importance and Similarity to degree of discrepancy.

- Incorporate Importance and Similarity scores into overall “Addition” “Deletion” and “Overlap” scores for the pair of reports.

Preliminary report
Final report

Each token in final report

$$S_a(\overline{p}, \overline{f}) = \sum_{f_i \in f}$$

Addition score

- We combine these scores using a multi-layer perceptron to produce a final ranking score.

Going from Importance and Similarity to degree of discrepancy.

- Incorporate Importance and Similarity scores into overall “Addition” “Deletion” and “Overlap” scores for the pair of reports.

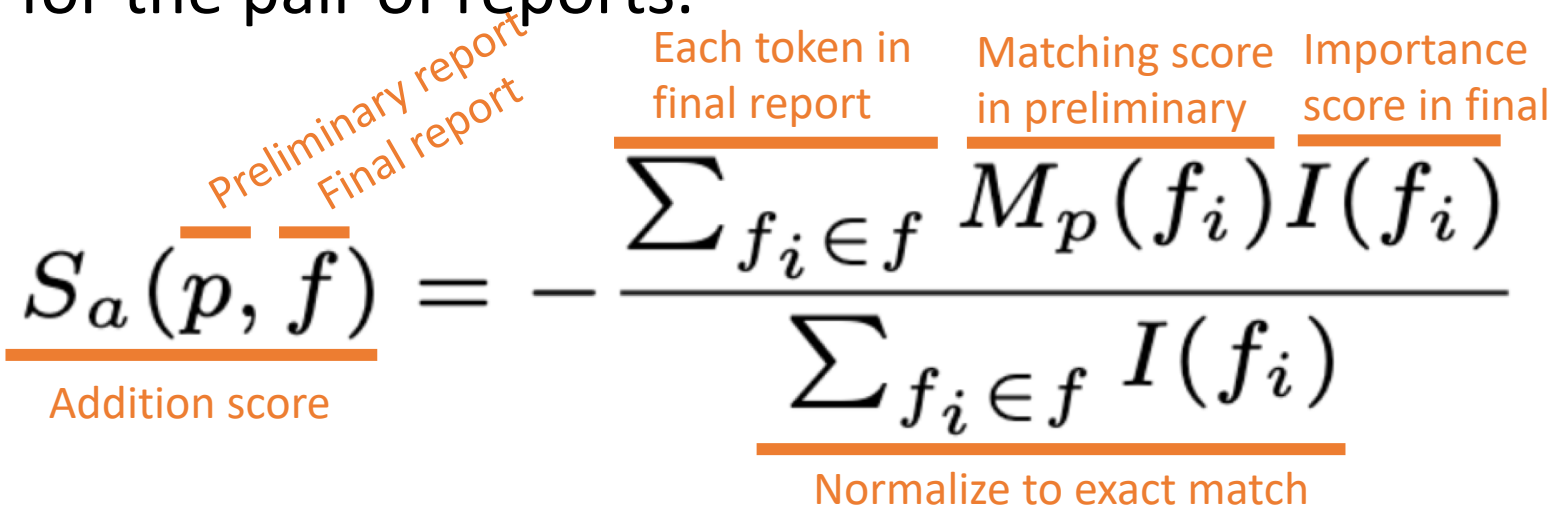
Preliminary report
Final report

$$\underbrace{S_a(\overline{p}, \overline{f})}_{\text{Addition score}} = \sum_{f_i \in f} \underbrace{M_p(f_i)}_{\substack{\text{Each token in} \\ \text{final report}}} \underbrace{I(f_i)}_{\substack{\text{Matching score} \\ \text{in preliminary}}} \underbrace{}_{\substack{\text{Importance} \\ \text{score in final}}}$$

- We combine these scores using a multi-layer perceptron to produce a final ranking score.

Going from Importance and Similarity to degree of discrepancy.

- Incorporate Importance and Similarity scores into overall “Addition” “Deletion” and “Overlap” scores for the pair of reports.



- We combine these scores using a multi-layer perceptron to produce a final ranking score.

Importance and Similarity Scores are defined over:

- Unigrams
  - SciBERT token representations
  - SciBERT is a transformer-based contextualized language model trained on
  - Importance score linear linear combination of above
- N-grams
  - Average embeddings over sliding SciBERT window
  - “Left arm” != “right arm”
  - Importance score linear combination of above
- Ontological matches
  - From RadLex
  - Synonym matches for similarity
    - Chauffeur’s fracture == Hutchinson fracture
  - Constant importance

We use data collected and annotated from MedStar

- Attending radiologists asked to annotate discrepancies on a 4-point scale after making their edits
  - Range from non-significant change (0) to obvious mistake (3)

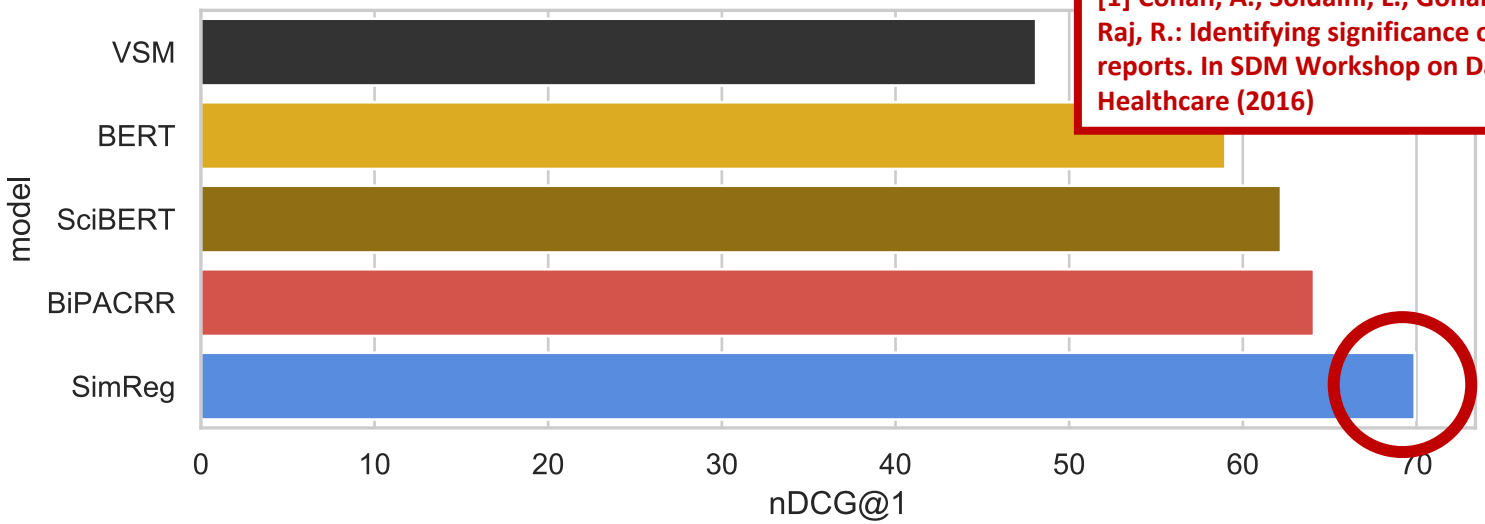
- 3,368 reports split into 122 sets based on resident-week sets to rank.
- 60/20/20 split by resident-week set.

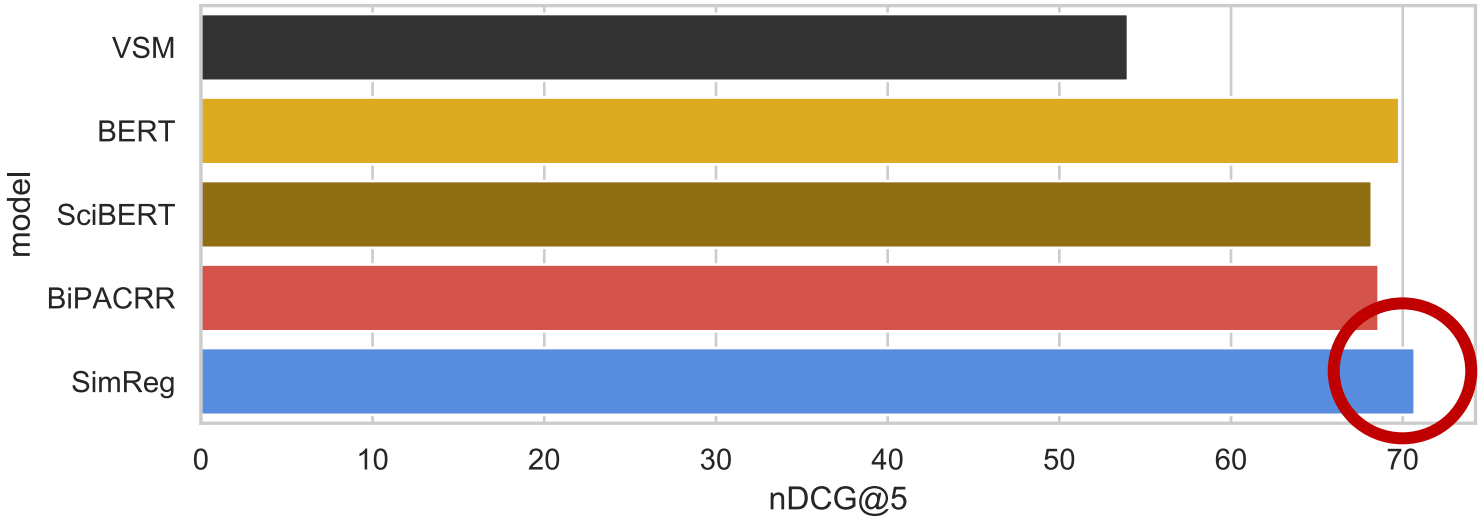
MedStar Health

Baseline methods

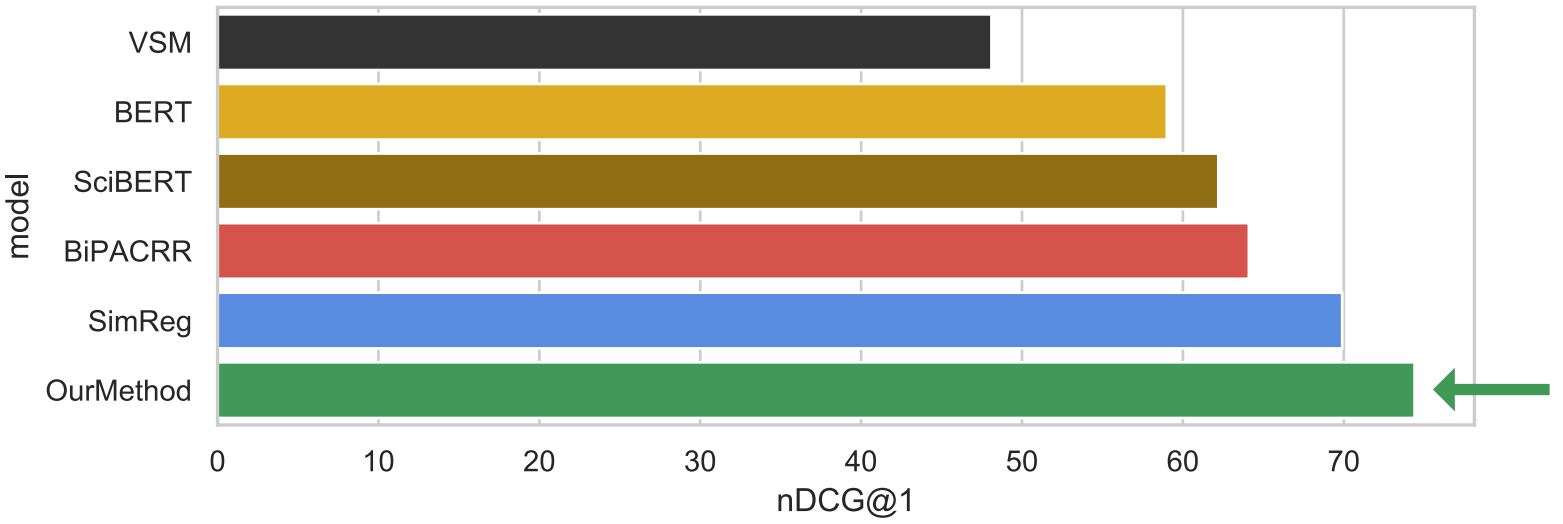
Feature-based SimReg [1] does well at finding reports with high degrees of discrepancy, but does not do as well ranking the rest of the list.

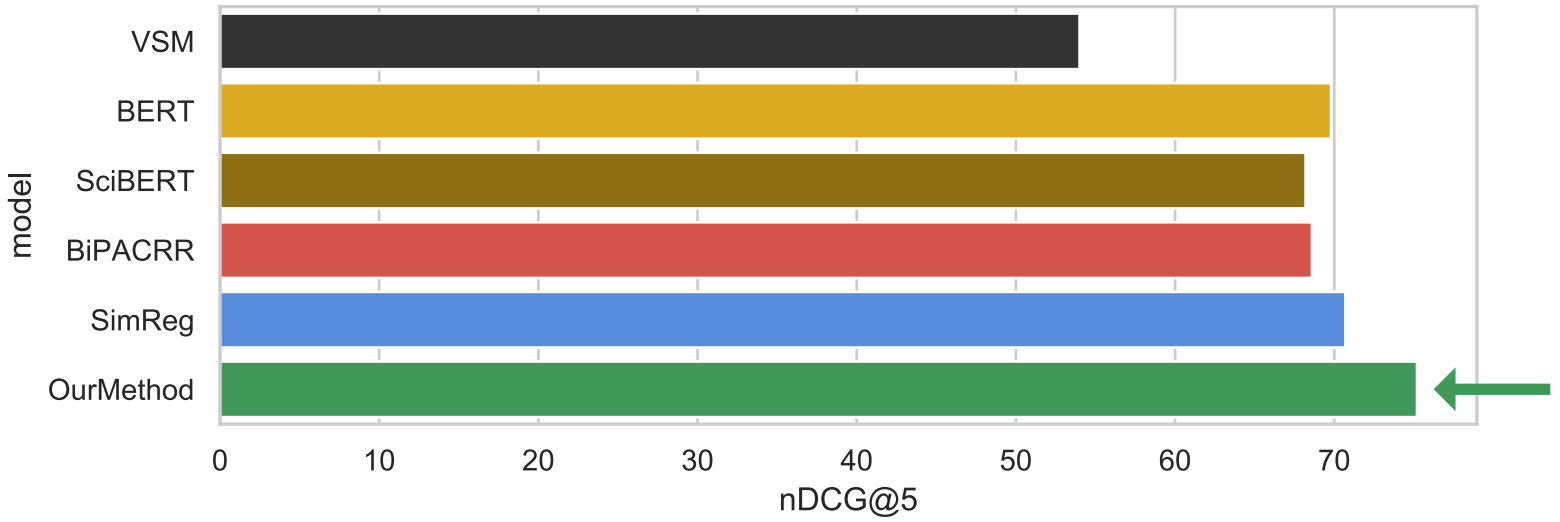
[1] Cohan, A., Soldaini, L., Goharian, N., Fong, A., Ross, F., Raj, R.: Identifying significance of discrepancies in radiology reports. In SDM Workshop on Data Mining for Medicine and Healthcare (2016)



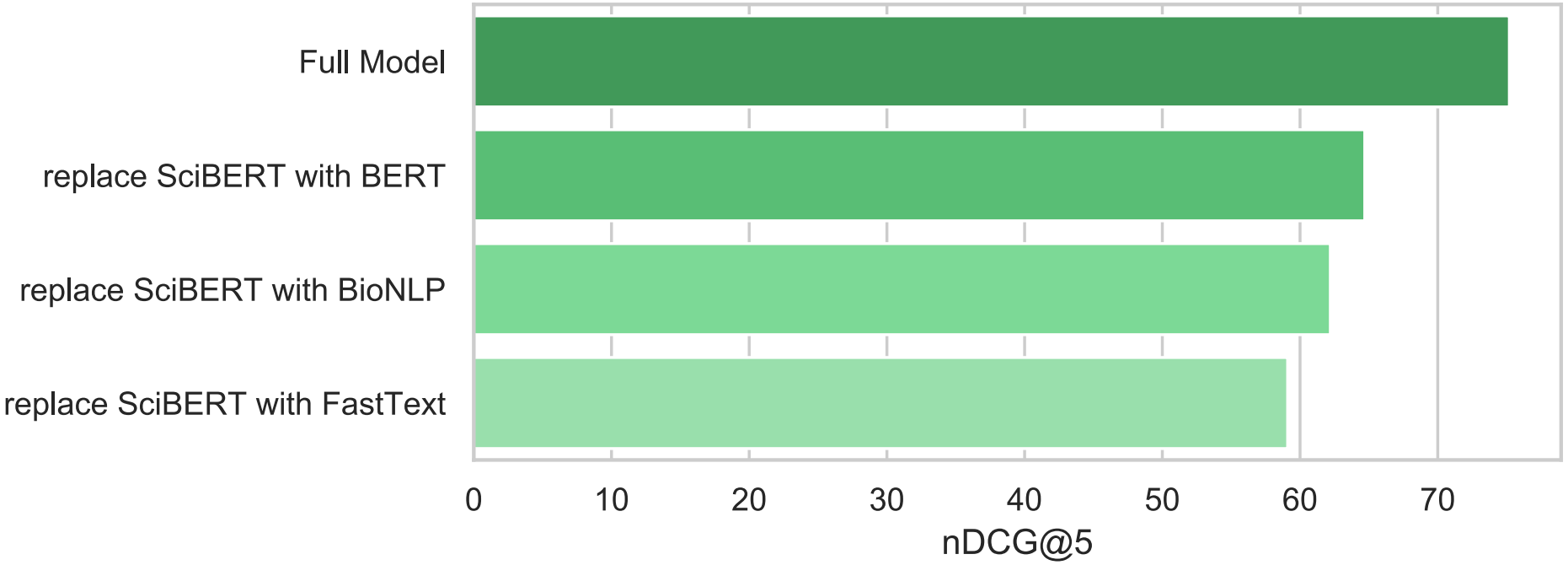


Our approach does better.

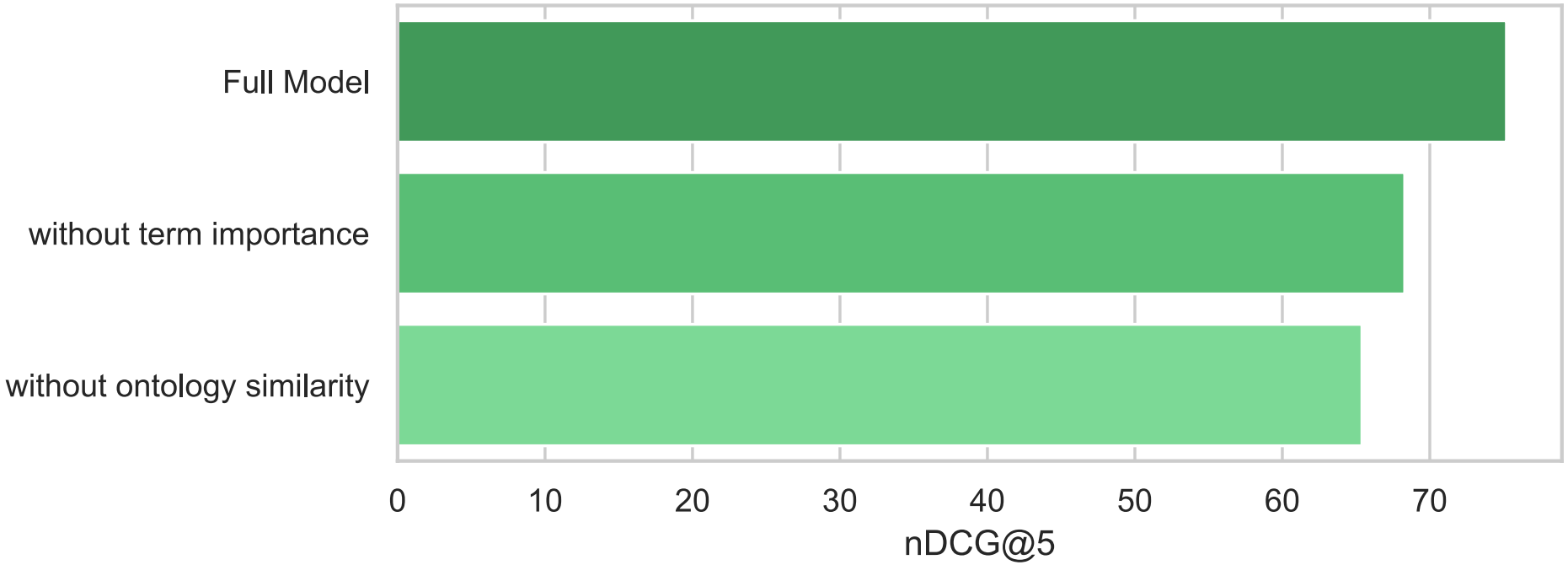




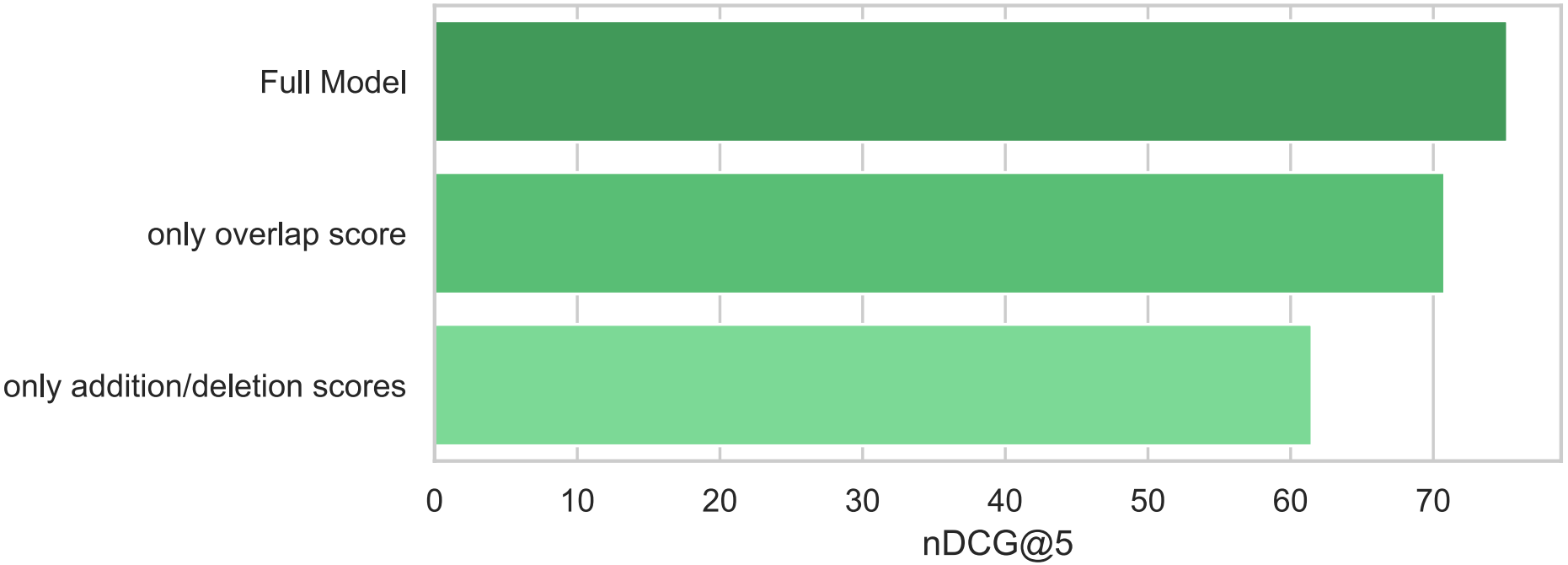
Ablation:
Pre-training contextualized model on biomedical texts is important.



Ablation:
Both importance and similarity scores play an important role.



Ablation:
The overlap score does most of the heavy lifting



Evaluating Term Importance

- Example excerpt:

~~anteroinferior~~ ~~dislocation~~ ~~of~~ ~~the~~ ~~left~~ ~~shoulder.~~ ~~mild~~ ~~hill-~~ ~~sachs~~ ~~deformity~~ ~~without~~ ~~associated~~ ~~bankart~~ ~~lesion.~~ no evidence of acute fracture or dislocation ~~of~~ ~~the~~ ~~humerus~~.

(darker colors higher importance scores)

- Words that are commonly given high importance:

no *cardiopulmonary* *abnormality*

Ranking Significant Discrepancies in Clinical Reports

Sean MacAvaney, Arman Cohan, Nazli Goharian, Ross Filice

- Finding discrepancies in radiology reports can be re-framed as a ranking problem.
- Term importance and similarity can be effective signals for ranking.
- The text representation matters a lot! Both contextualization and training in proper domain is important.

ir LAB GEORGETOWN UNIVERSITY

AI2

MedStar Health

Extra Slides

$$S_a(p, f) = -\frac{\sum_{f_i \in f} M_p(f_i) I(f_i)}{\sum_{f_i \in f} I(f_i)}$$

$$S_d(p, f) = -\frac{\sum_{p_i \in p} M_f(p_i) I(p_i)}{\sum_{p_i \in p} I(p_i)}$$

$$S_o(p, f) = -\frac{\sum_{p_i \in p} M_f(p_i) I(p_i) + \sum_{f_i \in f} M_p(f_i) I(f_i)}{\sum_{p_i \in p} I(p_i) + \sum_{f_i \in f} I(f_i)}$$

Dataset label distribution

	Ranking	Reports per discrepancy label			
Split	sets	0	1	2	3
Train	72	1,741	226	127	15
Dev	24	431	67	25	4
Test	26	557	108	57	10