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
Outline

• Motivation
• Framework
• Evaluation
• Conclusions
Overview of Clinical Workflow
### Example

<table>
<thead>
<tr>
<th>Initial Report</th>
<th>Final Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>No acute hemorrhage. No extra-axial fluid collections. The differentiation of gray and white matter is normal.</td>
<td>“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.”</td>
</tr>
</tbody>
</table>
Introduction

- Two types of discrepancies

<table>
<thead>
<tr>
<th>Preliminary report (resident radiologist)</th>
<th>Significant discrepancies</th>
<th>Non-significant discrepancy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>“No acute hemorrhage. No extra-axial fluid collections. The differentiation of gray and white matter is normal.”</td>
<td>“Postsurgical changes related to right thoracotomy with surgical packing material and hemorrhagic blood products in the right lower chest.”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Final report (attending radiologist)</th>
<th>Significant discrepancies</th>
<th>Non-significant discrepancy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>“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.”</td>
<td>“Postsurgical changes related to right thoracotomy with surgical packing material and large amount of hemorrhagic blood products in the right lower chest.”</td>
</tr>
</tbody>
</table>
# Introduction

- **Two types of discrepancies**

<table>
<thead>
<tr>
<th></th>
<th>Significant discrepancies</th>
<th>Non-significant discrepancy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Preliminary report</strong></td>
<td>“No acute hemorrhage. No extra-axial fluid collections. The differentiation of gray and white matter is normal.”</td>
<td>“Postsurgical changes related to right thoracotomy with surgical packing material and hemorrhagic blood products in the right lower chest.”</td>
</tr>
<tr>
<td>(resident radiologist)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Final report</strong></td>
<td>“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.”</td>
<td>“Postsurgical changes related to right thoracotomy with surgical packing material and large amount of hemorrhagic blood products in the right lower chest.”</td>
</tr>
<tr>
<td>(attending radiologist)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Introduction

- Two types of discrepancies

<table>
<thead>
<tr>
<th></th>
<th>Significant discrepancies</th>
<th>Non-significant discrepancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>**Preliminary</td>
<td>“No acute hemorrhage. No extra-axial fluid collections. The differentiation of gray and white matter is normal.”</td>
<td>“Postsurgical changes related to right thoracotomy with surgical packing material and hemorrhagic blood products in the right lower chest.”</td>
</tr>
<tr>
<td>report</td>
<td>(resident radiologist)</td>
<td></td>
</tr>
<tr>
<td>**Final report</td>
<td>“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.”</td>
<td>“Postsurgical changes related to right thoracotomy with surgical packing material and large amount of hemorrhagic blood products in the right lower chest.”</td>
</tr>
</tbody>
</table>
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
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

<table>
<thead>
<tr>
<th></th>
<th>Radlex</th>
<th>Human A</th>
<th>Human B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Significant</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Radlex</td>
<td>1.0</td>
<td>0.964</td>
<td>0.942</td>
</tr>
<tr>
<td>Human A</td>
<td>0.946</td>
<td>1.0</td>
<td>0.906</td>
</tr>
<tr>
<td>Human B</td>
<td>0.942</td>
<td>0.906</td>
<td>1.0</td>
</tr>
<tr>
<td>Count=139</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fleiss $\kappa = 0.880$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Stage-1: Radlex heuristic

<table>
<thead>
<tr>
<th>Non-Significant</th>
<th>Radlex</th>
<th>Human A</th>
<th>Human B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radlex</td>
<td>1.0</td>
<td>0.964</td>
<td>0.942</td>
</tr>
<tr>
<td>Human A</td>
<td>0.946</td>
<td>1.0</td>
<td>0.906</td>
</tr>
<tr>
<td>Human B</td>
<td>0.942</td>
<td>0.906</td>
<td>1.0</td>
</tr>
<tr>
<td>Count</td>
<td>139</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fleiss $\kappa$</td>
<td>0.880</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Significant</th>
<th>Radlex</th>
<th>Human A</th>
<th>Human B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radlex</td>
<td>1.0</td>
<td>0.557</td>
<td>0.492</td>
</tr>
<tr>
<td>Human A</td>
<td>0.557</td>
<td>1.0</td>
<td>0.934</td>
</tr>
<tr>
<td>Human B</td>
<td>0.492</td>
<td>0.934</td>
<td>1.0</td>
</tr>
<tr>
<td>Count</td>
<td>61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fleiss $\kappa$</td>
<td>0.468</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

- Baseline

Graph showing the AUC for F-1 and AUC with baseline values.
Results: Individual Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Baseline</th>
<th>RADLEX</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-1</td>
<td>0.64</td>
<td>0.68</td>
</tr>
<tr>
<td>AUC</td>
<td>0.58</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Identifying Significance of Discrepancies in Radiology Reports (SDM-DMMH 16)
Results: Individual Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Baseline</th>
<th>RADLEX</th>
<th>BLEU</th>
<th>ROUGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-1</td>
<td>0.65</td>
<td>0.7</td>
<td>0.7</td>
<td>0.75</td>
</tr>
<tr>
<td>AUC</td>
<td>0.7</td>
<td>0.8</td>
<td>0.8</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Identifying Significance of Discrepancies in Radiology Reports (SDM-DMMH 16)
Results: Individual Features

- F-1
- AUC

Baselines:
- Baseline
- RADLEX
- BLEU
- ROUGE

Identifying Significance of Discrepancies in Radiology Reports (SDM-DMMH 16)
Results: Individual Features

Identifying Significance of Discrepancies in Radiology Reports (SDM-DMMH 16)
Feature Combinations
Summarization and MT

Baseline    ROUGE + BLEU    ROUGE + METEOR    ROUGE + WER

![Graph showing F-1 and AUC scores for different feature combinations.](image)
Feature Combinations

- Baseline
- Rg+BL
- Rg+WER
- Rg+BL+M
- Rg+BL+M+RL
- Rg+BL+M+RL+Rd

Graph showing feature combinations and their respective F-1 and ACC scores.
Feature Combinations

- Baseline
- ROUGE + BLEU
- ROUGE + METEOR
- ROUGE + BLEU + METEOR
- RG + BL + MET + RAD
- RG + BL + MET + RAD + READ

Bar chart showing F-1 and AUC for different feature combinations.
Feature Combinations

- Baseline
- ROUGE + BLEU
- ROUGE + METEOR
- ROUGE + BLEU + METEOR
- RG + BL + MET + RAD
- RG + BL + MET + RAD + READ

Graph showing the performance of different feature combinations for F-1 and AUC metrics.
Feature Combinations

- Baseline
- ROUGE + BLEU
- ROUGE + METEOR
- ROUGE + BLEU + METEOR
- RG+BL+MET+RAD
- RG+BL+MET+RAD+READ

Bar chart showing performance metrics (F-1 and AUC) for different feature combinations.
Feature Combinations

- Baseline
- ROUGE + BLEU
- ROUGE + METEOR
- RG+BL+MET+RAD
- RG+BL+MET+RAD+READ

Bar chart comparing performance metrics (F-1 and AUC) for different feature combinations.
Feature Combinations

Baseline

ROUGE + BLEU

ROUGE + METEOR

ROUGE+BLEU+METEOR

RG+BL+MET+RAD

RG+BL+MET+RAD+READ

<table>
<thead>
<tr>
<th>F-1</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>0.85</td>
<td></td>
</tr>
</tbody>
</table>
False Negative Rate

- Baseline
- ROUGE + BLEU
- ROUGE + METEOR
- ROUGE + WER
- ROUGE+BLEU+METEOR
- RG+BL+MET+RAD+READ

FNR

0.1
0.12
0.14
0.16
0.18
0.2
0.22
0.24
0.26
0.28
0.3
0.32
0.34
0.36
Proposed Features vs Baseline

Receiver Operating Characteristic

- **PROPOSED FEATURES** (area = 0.837)
- **RADLEX** (area = 0.730)
- **RADLEX+SURFACE** (area = 0.746)
Feature Comparison

Receiver Operating Characteristic

- BLEU (area = 0.757)
- METEOR (area = 0.627)
- RADLEX (area = 0.746)
- READABILITY (area = 0.594)
- ROUGE (area = 0.838)
- SURFACE (area = 0.643)
- WER (area = 0.704)
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