

Guidance in Radiology Report Summarization: An Empirical Evaluation and Error Analysis

Jan Trienes, Paul Youssef, Jörg Schlötterer, Christin Seifert

INLG 2023, Prague



Universitätsmedizin Essen
Universitätsklinikum



Open-Minded

Philipps



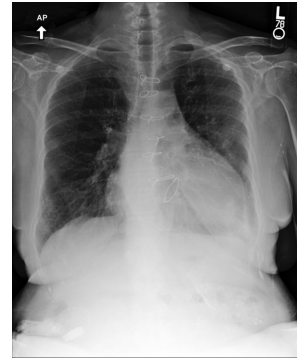
Universität
Marburg

Task: automatic impression generation

Background: Technique: Chest, AP and lateral. Comparison: _ and _. History: Weakness and decreased blood sugar with leg swelling and tenderness.

Findings: The patient is status post coronary artery bypass graft surgery and apparently mitral valve replacement. The heart is mildly enlarged. The mediastinal and hilar contours appear unchanged. There is a slight interstitial abnormality, suggestive of a state of very mild congestion, but no new focal opacity. A left-sided pleural effusion has resolved although mild scarring or atelectasis persists. Bones are probably demineralized.

Impression: Findings suggesting mild pulmonary congestion. Resolution of small left-side pleural effusion.



Task: automatic impression generation

Background: Technique: Chest, AP and lateral. Comparison: _ and _. History: Weakness and decreased blood sugar with leg swelling and tenderness.

Findings: The patient is status post coronary artery bypass graft surgery and apparently mitral valve replacement. The heart is mildly enlarged. The mediastinal and hilar contours appear unchanged. There is a slight interstitial abnormality, suggestive of a state of very mild congestion, but no new focal opacity. A left-sided pleural effusion has resolved although mild scarring or atelectasis persists. Bones are probably demineralized.

Impression: Findings suggesting mild pulmonary congestion. Resolution of small left-side pleural effusion.

Information about the patients' condition and the procedure.

Detailed description of imaging observations. Positive and negative findings.

Summary of the most important observations. Typically 1/3 of the findings length.

Task: automatic impression generation

Background: Technique: Chest, AP and lateral.
Comparison: _ and _. History: Weakness and decreased blood sugar with leg swelling and tenderness.

Findings: The patient is status post coronary artery bypass graft surgery and apparently mitral valve replacement. The heart is mildly enlarged. The mediastinal and hilar contours appear unchanged. There is a slight interstitial abnormality, suggestive of a state of very mild congestion, but no new focal opacity. A left-sided pleural effusion has resolved although mild scarring or atelectasis persists. Bones are probably demineralized.

Impression: Findings suggesting mild pulmonary congestion. Resolution of small left-side pleural effusion.

Writing the impression is both extractive (selecting findings), and abstractive (forming a concise conclusion)

Research contributions

1

Abstractive summarization is difficult to control, prone to hallucination
Proposal: extractive summaries as guidance (cheap, domain-agnostic)

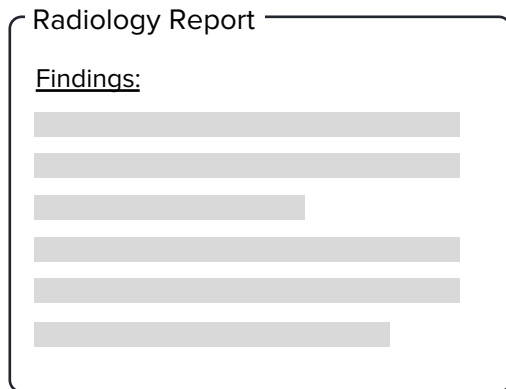
2

Automatic eval (ROUGE, ...) suggests progress. What problems remain?
Manual error analysis of (un)guided methods

Guided summarization framework

1. Extractive summarization (k sents.)

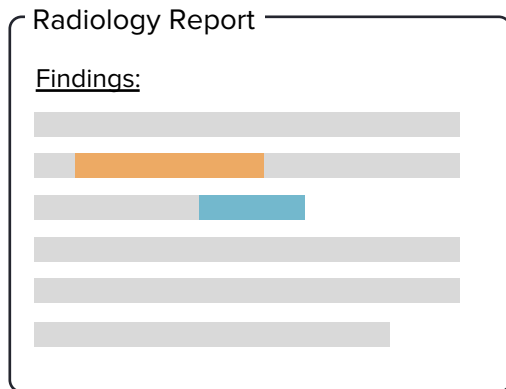
e.g., BertExt



Guided summarization framework

1. Extractive summarization (k sents.)

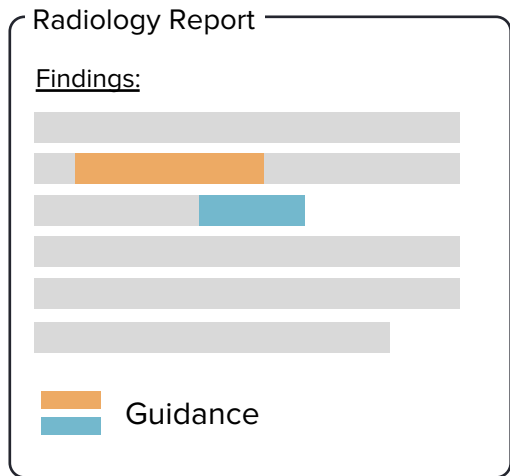
e.g., BertExt



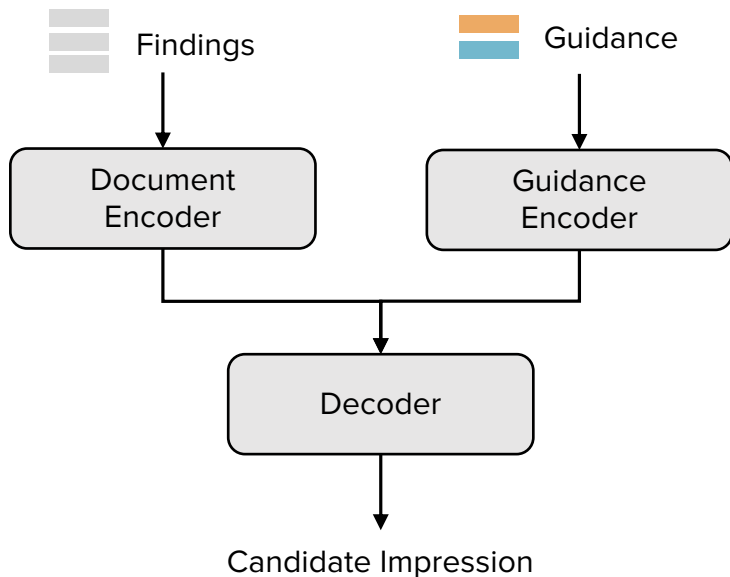
Guided summarization framework

1. Extractive summarization (k sents.)

e.g., BertExt

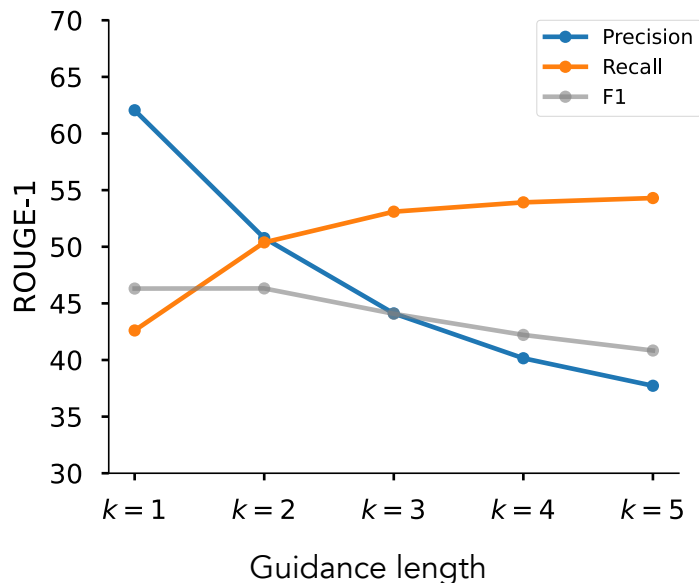


2. Guided abstractive summarization

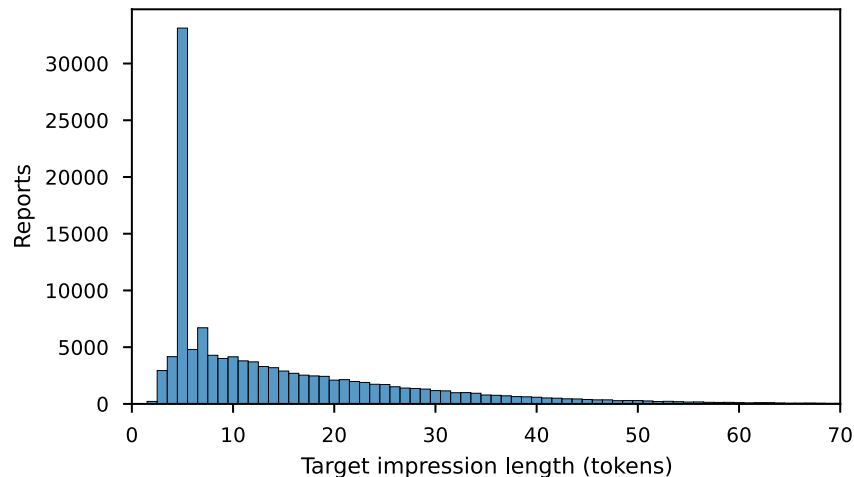


Fixed-length guidance signal for all reports is ineffective

Longer guidance improves recall,
but deteriorates overall quality



Hypothesis: effective guidance depends on
intended target length



Variable-length extractive summaries as guidance

Fixed-length extractive (Liu & Lapata, 2019)

- Obtain training labels from oracle
- Train binary sentence-level classifier
- Pick top-k sentences (for all docs)

Method 1: thresholding (ours)

- Pick all sents with $\geq T$, rather than top-k
- Learn T from val. set

Method 2: oracle approximation (ours)

- Learn classifier $f(x) \rightarrow k$
- Training labels is ROUGE Oracle

Fixed

s_i	ROUGE Oracle	BertExt $p(y = 1 s_i)$	Top-k (k=3)
1		0.1	
2	x	0.9	x
3	x	0.4	x
4		0.1	
5		0.2	x

Variable

Thresholding (e.g., $T \geq 0.4$)	Oracle Approx
x	x
x	x

Example doc (5 sentences)

Evaluation on two real world radiology datasets

Chest x-rays

	MIMIC-CXR	OpenI
Instances	122,500 / 963 / 1,598	2,342 / 334 / 670
Avg. Finding length	56 tokens	37 tokens
Avg. Impression length	15 tokens	8 tokens
Novelty (unigram)	73.4%	86.8%

Method	MIMIC-CXR					OpenI				
	R-1	R-2	R-L	BS	Fact.	R-1	R-2	R-L	BS	Fact.
<i>Baselines and fixed-length guidance</i>										
OracleExt										
BertExt (Liu and Lapata, 2019)										
BertAbs (Liu and Lapata, 2019)										
GSum (Dou et al., 2021)										
<i>Variable-length guidance (ours)</i>										
GSum w/ LR-Approx										
GSum w/ BERT-Approx										
GSum w/ Thresholding										
<i>Domain-specific methods</i>										
WGSum (Hu et al., 2021)										
WGSum+CL (Hu et al., 2022)										

Method	MIMIC-CXR					OpenI				
	R-1	R-2	R-L	BS	Fact.	R-1	R-2	R-L	BS	Fact.
<i>Baselines and fixed-length guidance</i>										
OracleExt	44.0	25.4	40.6	50.1	55.1					
BertExt (Liu and Lapata, 2019)	32.7	18.1	30.0	41.9	44.5					
BertAbs (Liu and Lapata, 2019)	48.4	34.1	46.6	58.8	47.3					
GSum (Dou et al., 2021)	46.3	32.7	44.7	57.4	46.6					
<i>Variable-length guidance (ours)</i>										
GSum w/ LR-Approx										
GSum w/ BERT-Approx										
GSum w/ Thresholding										
<i>Domain-specific methods</i>										
WGSum (Hu et al., 2021)										
WGSum+CL (Hu et al., 2022)										

Takeaways

1. Fixed-length guidance is worse than no guidance

Method	MIMIC-CXR					OpenI				
	R-1	R-2	R-L	BS	Fact.	R-1	R-2	R-L	BS	Fact.
<i>Baselines and fixed-length guidance</i>										
OracleExt	44.0	25.4	40.6	50.1	55.1					
BertExt (Liu and Lapata, 2019)	32.7	18.1	30.0	41.9	44.5					
BertAbs (Liu and Lapata, 2019)	48.4	34.1	46.6	58.8	47.3					
GSum (Dou et al., 2021)	46.3	32.7	44.7	57.4	46.6					
<i>Variable-length guidance (ours)</i>										
GSum w/ LR-Approx	48.9	34.2	47.0	59.1	48.2					
GSum w/ BERT-Approx	49.4	34.5	47.4	59.5	50.6					
GSum w/ Thresholding	49.9	34.3	47.8	59.8	49.0					
<i>Domain-specific methods</i>										
WGSum (Hu et al., 2021)	48.4	32.8	46.5	58.6	49.8					
WGSum+CL (Hu et al., 2022)	49.5	35.3	47.8	59.5	51.1					

Takeaways

1. Fixed-length guidance is worse than no guidance
2. Variable-length improves over unguided, competitive w/ domain-specific

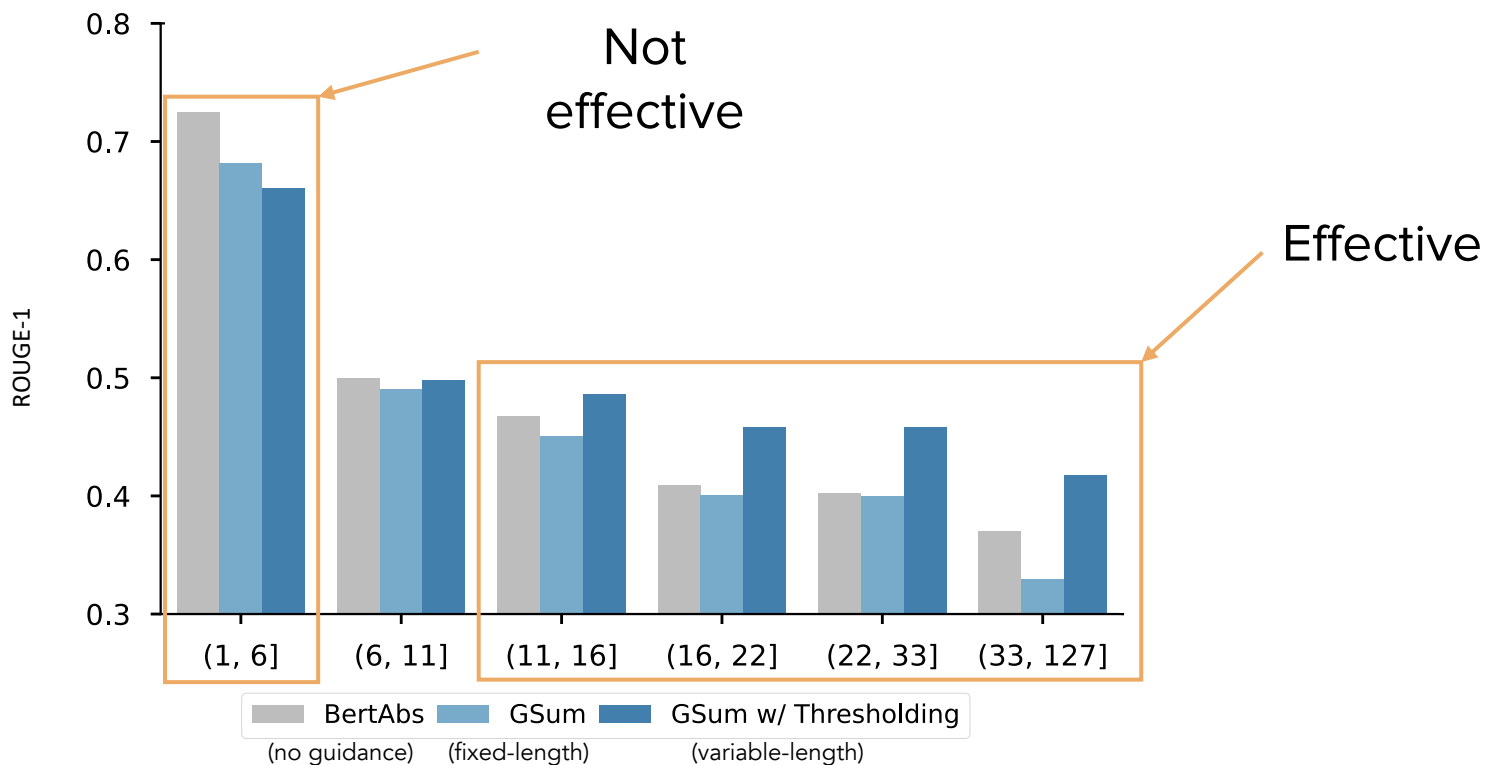
Method	MIMIC-CXR					OpenI				
	R-1	R-2	R-L	BS	Fact.	R-1	R-2	R-L	BS	Fact.
<i>Baselines and fixed-length guidance</i>										
OracleExt	44.0	25.4	40.6	50.1	55.1	30.5	11.9	29.2	33.7	53.5
BertExt (Liu and Lapata, 2019)	32.7	18.1	30.0	41.9	44.5	23.6	7.4	22.6	32.2	42.8
BertAbs (Liu and Lapata, 2019)	48.4	34.1	46.6	58.8	47.3	62.0	52.7	61.7	69.2	39.3
GSum (Dou et al., 2021)	46.3	32.7	44.7	57.4	46.6	60.1	49.6	59.8	67.0	40.0
<i>Variable-length guidance (ours)</i>										
GSum w/ LR-Approx	48.9	34.2	47.0	59.1	48.2	62.0	51.2	61.6	67.9	41.7
GSum w/ BERT-Approx	49.4	34.5	47.4	59.5	50.6	62.5	51.6	62.2	68.4	39.6
GSum w/ Thresholding	49.9	34.3	47.8	59.8	49.0	62.2	50.8	61.8	68.6	40.4
<i>Domain-specific methods</i>										
WGSum (Hu et al., 2021)	48.4	32.8	46.5	58.6	49.8	61.1	50.0	60.8	67.9	38.4
WGSum+CL (Hu et al., 2022)	49.5	35.3	47.8	59.5	51.1	64.7	57.1	64.5	70.0	37.2

Takeaways

1. Fixed-length guidance is worse than no guidance
2. Variable-length improves over unguided, competitive w/ domain-specific
3. On more abstractive data (OpenI) no clear benefit

Extractive guidance helps to generate longer summaries

ROUGE by target length (tokens)



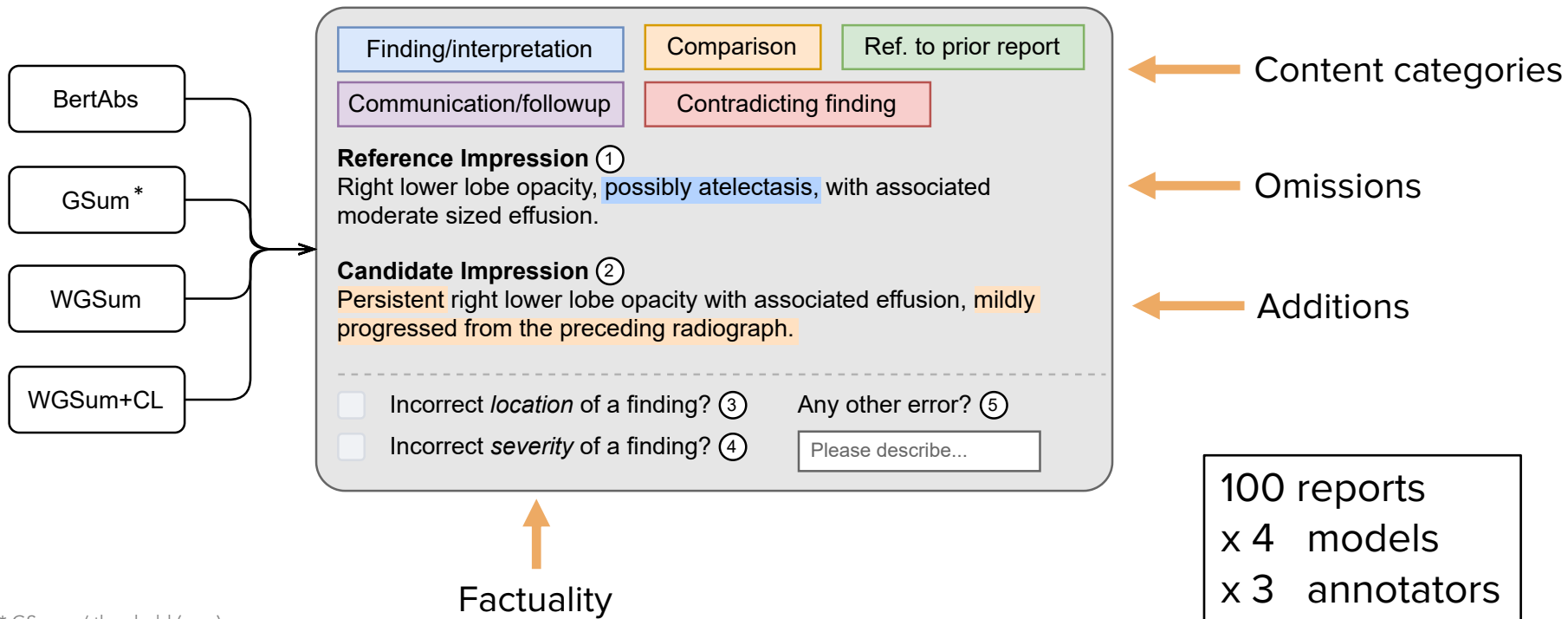
So far...

Guided methods are effective according to automatic evaluation.

What problems remain?

Manual error analysis of (un)guided methods

Error analysis protocol (inspired by MQM)



* GSum w/ threshold (ours)

Extended from the taxonomy of Yu et al. (2022)

Most frequent errors are addition/omission of findings

Guided methods reduce risk of omissions

Category: “findings”	Omissions	Additions
BertAbs (unguided)	70	51
GSum w/ threshold (ours)	58	72
WGSum (Hu et al., 2021)	62	61
WGSum+CL (Hu et al., 2022)	64	<u>54</u>

Reference: Interval increase in vascular engorgement. No frank interstitial edema. **No focal consolidations identified.**

Candidate: interval increase in pulmonary vascular congestion without evidence of interstitial edema. **small right-sided pleural effusion.**

Majority of overlapping findings is factual!

- Incorrect location (5-8%)
- Incorrect severity (7-9%)

Reference: ... there is **near-complete** resolution of pleural effusion

Candidate: ... there is resolution of pleural effusion

Target impressions sometimes contain **followups**

Cannot be generated without clinical context

Category: “followups”	Omissions	Additions
BertAbs (unguided)	20	5
GSum w/ threshold (ours)	18	8
WGSum (Hu et al., 2021)	19	8
WGSum+CL (Hu et al., 2022)	19	4

Reference: Multiloculated right pleural effusion unchanged since __. [...] Findings were relayed to Dr. _ by Dr. __ following review on _ at 11:00 via telephone.

Candidate: stable appearance of multiple loculated right pleural effusion.

Hallucinations!



See the paper for all 11 categories and more examples...

#	Error Category	M1 (%)	M2 (%)	M3 (%)	M4 (%)
0	No error	20 (20)	18 (18)	14 (14)	22 (22)
<i>Omissions from reference</i>					
1a	Finding/interpretation	70 (52)	58 (43)	62 (48)	64 (47)
1b	Comparison	23 (19)	16 (15)	19 (16)	23 (19)
1c	Ref. to prior report	1 (1)	3 (3)	2 (2)	2 (2)
1d	Communication/followup	20 (19)	18 (16)	19 (17)	19 (17)
Total		114 (66)	95 (58)	102 (63)	108 (61)
<i>Additions to candidate</i>					
2a	Finding/interpretation	51 (44)	72 (57)	61 (50)	54 (46)
2b	Comparison	11 (8)	10 (9)	9 (9)	7 (6)
2c	Ref. to prior report	0 (0)	1 (1)	0 (0)	0 (0)
2d	Communication/followup	5 (5)	8 (6)	8 (8)	4 (3)
2e	Contradicting finding	0 (0)	1 (1)	3 (3)	1 (1)
Total		67 (49)	92 (63)	81 (58)	66 (48)
<i>Semantics of intersecting findings</i>					
3	Incorrect location	5 (5)	8 (8)	8 (8)	7 (7)
4	Incorrect severity	6 (6)	7 (7)	7 (7)	9 (9)
5	Other error	31 (23)	30 (23)	33 (29)	30 (21)

Reference: Interval increase in vascular engorgement. No frank interstitial edema. **No focal consolidations identified.**

Candidate (M3): interval increase in pulmonary vascular congestion without evidence of interstitial edema. **small right-sided pleural effusion.**

Reference: Right lower lobe opacity, **possibly atelectasis**, with associated moderate sized effusion.

Candidate (M4): **persistent** right lower lobe opacity with associated effusion, **mildly progressed from the preceding radiograph.**

Reference: Multiloculated right pleural effusion unchanged **since _**. **New linear and nodular opacities in the left upper lobe may represent carcinomatosis.** Findings were relayed to Dr. _ by Dr. _ following review on _ at approximately 11:00 via telephone.

Candidate (M1): **stable appearance of multiple loculated right pleural effusion.**

Reference: Unchanged size and position of right-sided hydropneumothorax over the last _-hour examination interval.

Candidate (M3): **development of new right-sided hydropneumothorax** in this patient with history of newly placed pigtail catheter. **referring physician, _ was paged at 4:45 p.m.**

Reference: Little change in the severe bronchiectasis and **emphysema.**

Candidate (M3): **unchanged** bibasilar bronchiectasis and bibasilar bronchiectasis.

Figure 4: Results of manual error analysis of 100 MIMIC-CXR reports. Left: number of times each error occurred per method (percent of reports in gray, least errors per row in bold). Right: example error annotations. Models: BertAbs (M1), GSum w/ Thresholding (M2), WGSum (M3), and WGSum+CL (M4) [best viewed in color].

How to explain the problems in content selection?

Our ✨ hypothesis: ✨ models lack clinical context

Latent factors in reporting

- Patient demographics
- Chest x-ray (multimodal)
- What happened after?
- What happened before?

Background section

Background: Technique: Chest, AP and lateral. Comparison: _ and _. History: Weakness and decreased blood sugar with leg swelling and tenderness.

Findings: The patient is status post coronary artery bypass graft surgery and apparently mitral valve replacement. The heart is mildly enlarged. The mediastinal and hilar contours appear unchanged. There is a slight interstitial abnormality, suggestive of a state of very mild congestion, but no new focal opacity. A left-sided pleural effusion has resolved although mild scarring or atelectasis persists. Bones are probably demineralized.

Impression: Findings suggesting mild pulmonary congestion. Resolution of small left-side pleural effusion.

How to explain the problems in content selection?

Our ✨ hypothesis: ✨ models lack clinical context

Latent factors in reporting

- Patient demographics
- Chest x-ray (multimodal)
- What happened before?
- What happened after?

Background section

Almost all methods benefit from including background

	MIMIC-CXR					OpenI				
	R-1	R-2	R-L	BS	Fact.	R-1	R-2	R-L	BS	Fact.
OracleExt	0.2	-0.1	0.3	0.3	0.6	0	-0.4	-0.1	0.4	-1.8
BertExt	-0.2	0.1	0.2	0.2	0.2	-0.5	-0.6	-0.6	-1.1	-3.8
BertAbs	1.5	1.2	1.5	1.4	4.5	4.3	5.5	4.3	2.2	1.9
GSum (fixed)	2	1.6	1.7	1.8	3.3	2.7	3.6	2.7	2.4	0.4
GSum w/ LR-Approx	1.6	1.4	1.4	1.5	2	1.2	1.5	1.3	2.1	-1.8
GSum w/ BERT-Approx	1.5	1.3	1.4	1.3	0.1	1	1.2	1.1	1.8	0.2
GSum w/ Thresholding	1.5	1.7	1.6	1.5	3	2.1	3	2.1	2.1	1
WGSum	2.2	2.1	2	1.9	3	-2.5	-3	-2.3	-1.3	-1.6
WGSum+CL	2	1.9	2	2	-1.1	0.5	-0.5	0.4	1.1	5.9

(delta over training without background)

Conclusion


Takeaways


- Guided methods **effective at steering content selection**
- Extractive summaries are useful guidance, if we **adapt length to each document**
- **Content selection issues remain** – latent factors can explain some choices

Future work

- Incorporate more clinical context (multimodal, EHR data, clinician in the loop?)
- Benchmark and improve automatic metrics
 - **We release our error annotations**

Thanks!

 github.com/jantrienes/inlg2023-radsum

 jan.trienes@uni-due.de
paul.youssef@uni-marburg.de

Extra slides

Limitations of error analysis

Reference-based evaluation

Reference is most **reliable**
benchmark for importance (w/o
knowledge of clinical context)

Are additions faithful to findings?

Check **addition** spans for
entailment with all input sentences.

Model	Entail	Neutral	Contradict
BertAbs	31.9%	44.7%	23.4%
GSum w/ Thresholding	34.5%	36.2%	29.3%
WGSum	32.0%	44.0%	24.0%
WGSum+CL	33.3%	41.2%	25.5%

Preliminary

- Room for improvement in factuality
- Include findings in future annotations

Comparisons to prior studies also often added/omitted

Similar trend across all methods

Category: "comparisons"	Omissions	Additions
BertAbs (unguided)	23	11
GSum w/ threshold (ours)	16	10
WGSum (Hu et al., 2021)	19	9
WGSum+CL (Hu et al., 2022)	23	7

Reference: Right lower lobe opacity, possibly atelectasis, with associated moderate sized effusion.

Candidate: persistent right lower lobe opacity with associated effusion, mildly progressed from the preceding radiograph.

ROUGE oracle

Algorithm 1 Greedy Selection Algorithm

Input: A source document x consisting of multiple sentences $\{x_1, \dots, x_{|x|}\}$, its reference summary y , and a pre-defined integer N

Output: Oracle-selected highlighted sentences o

$o = \{\}$

for $i = 1, \dots, N$ **do**

$\text{max_rouge} = 0$

for s in x/o **do**

$\text{rouge}_1, \text{rouge}_2 = \text{cal_rouge}(o \cup \{s\})$

$\text{cur_rouge} = \text{rouge}_1 + \text{rouge}_2$

if $\text{cur_rouge} > \text{max_rouge}$ **then**

$\text{max_rouge} = \text{cur_rouge}$

$\text{max_sent} = s$

end if

end for

if $\text{max_rouge} == 0$ **then**

break

end if

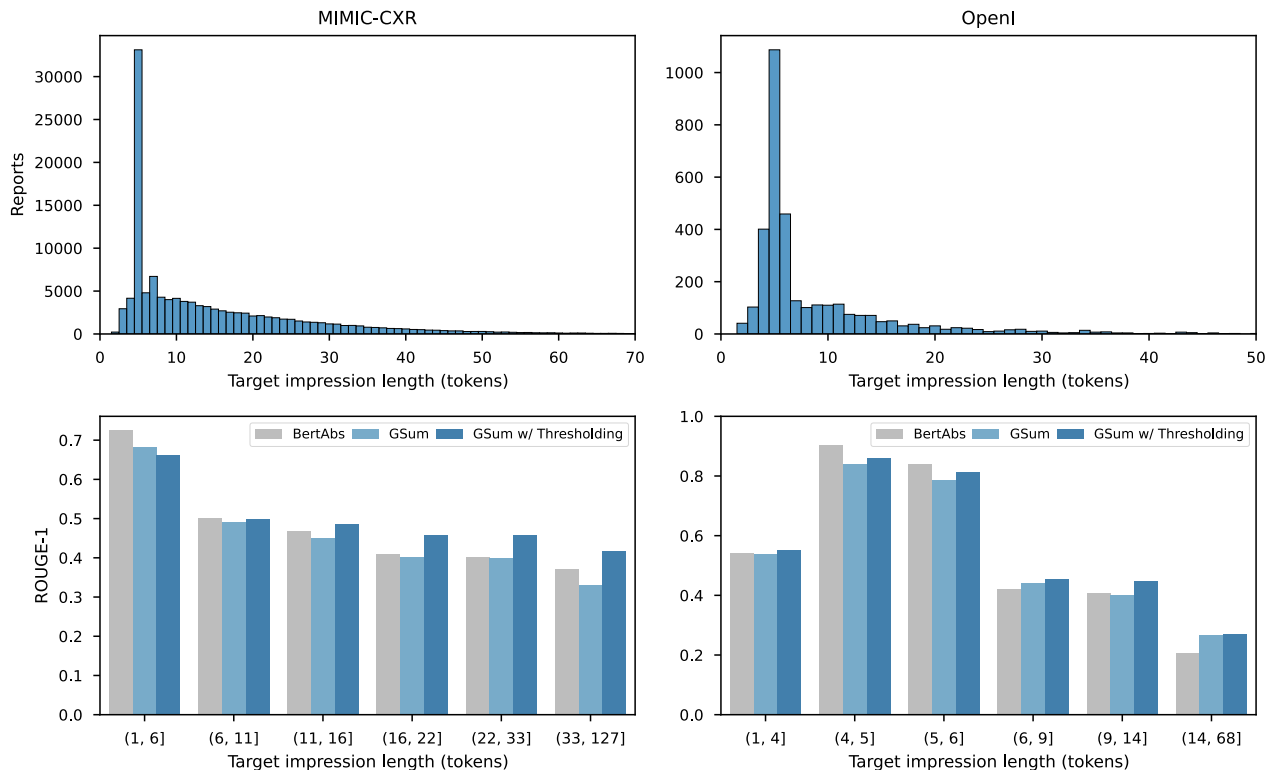
$o = o \cup \{ \text{max_sent} \}$

end for

return o

Extractive guidance helps to generate longer summaries

ROUGE by target length (tokens)



Aggregating span-based annotation

First align, then majority vote

Tokens:	a	b	c	d	e	f	g	h
A1	:	[-e1-]		[-e2-]				
A2	:	[-e1-]	[-e1-]		[-e2-]			
A3	:	[-e1-]					[-e1-]	

Group	:	1		2			3	

Vote	:	[-e1-]		[-e2-]				

Inter-annotator agreement

F1 for span-based (1, 2)

Krippendorff's alpha (3, 4)

#	Category	IAA	Count
<i>Omissions from reference</i>			
1a	Finding/interpretation	0.64	774
1b	Comparison	0.34	236
1c	Ref. to prior report	0.23	43
1d	Communication/followup	0.83	216
Total		0.61	1269
<i>Additions to candidate</i>			
2a	Finding/interpretation	0.66	718
2b	Comparison	0.44	155
2c	Ref. to prior report	0.08	17
2d	Communication/followup	0.65	72
2e	Contradicting finding	0.26	34
Total		0.60	996
3	Incorrect location	0.26	111
4	Incorrect severity	0.41	121

Table 11: Inter-annotator agreement (IAA) by category and total number of annotations before majority voting.