A System for Automatic Summarization and Simplification of Radiology Reports

Keen You,[1] Sophie Chheang,[2] Ali Mozayan,[2] Pratheek Bobba,[2] Spencer Beck,[2] Anne Sailer,[2] Muzz Muhammad,[2] August Allocco,[2] Irene Li, [1] and Dragomir Radev [1]

[1]Department of Computer Science, Yale University, New Haven, CT and [2]Yale School of Medicine, New Haven, CT

LILY Lab

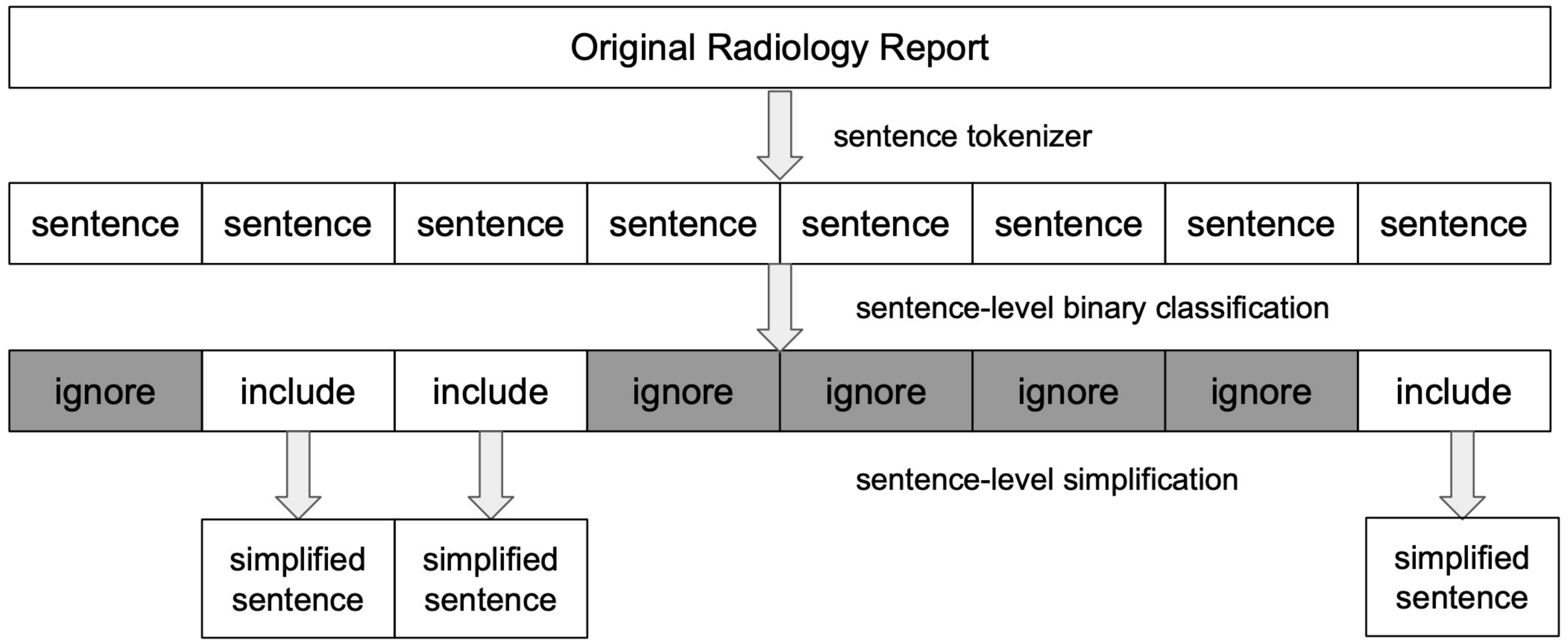
Introduction

Advances in NLP research have been increasingly applied to the medical domain since ~80% of Electronic Health Records (EHR) exists in textual form. One such application is to use text simplification tools on medical documents to make them to more accessible to the general public. It is useful as patients who have access to their health records have more sense of control over their health conditions and hence more likely to take medications and follow professional advice. However, the effectiveness of providing EHR to patients is greatly reduced by the difficulty of understanding such documents due to the extensive use of professional terminologies. Among all EHR, radiology reports are the most difficult to understand. Previous works in radiology report simplification utilize lexicons or knowledge bases to substitute difficult terms that are present. This approach has shown limited effectiveness due to lack of coverage and inadequate simplification. In this work, we investigate the simplification of radiology report by proposing a new task formulation and an auxiliary annotated dataset of 28 reports.

Task & Data

We propose a two-phase task formulation, the overview is illustrated in Figure 1. In the pre-processing step, each radiology report is tokenized into individual sentences. In the first phase, a binary classifier identifies each sentence as *include* or *ignore* in the simplified output. In the second phase, each sentence that is classified as *include* is transformed into a simplified version.

A total of 189 de-identified full radiology reports with a focus on chest CTs are provided by Yale School of Medicine. Metadata of the 189 reports are provided in Table 1. Among these reports, 28 reports of varying lengths are selected as our first batch of annotation. Metadata for the 28 reports are reported in Table 2. During annotation, every sentence is given a binary label, indicating whether it should be discarded. For every sentence that is to be kept, a radiology expert provides a gold simplification sequence of text.



Original Radiology Report

sentence tokenizer

sentence | sentence | sentence | sentence | sentence | sentence | sentence | sentence

sentence-level binary classification

ignore | include | include | ignore | ignore | ignore | ignore | include

sentence-level simplification

simplified sentence | simplified sentence | simplified sentence

Figure 1. Overview of task formulation.

avg # of sentence per report	15.21
total number of sentences	2875
avg # of tokens per report	158.84
avg # of tokens per sentence	10.44

Table 1. Metadata for the 189 unannotated reports.

# of annotated sentences	451
avg # of input tokens	10.29
avg # of simplified tokens	9.76
# ignored sentences	311
# included sentences	140

Table 2 . Metadata for annotated 28 reports. With ~70% of sentences labeled as *ignore*, writing simplified sentences for *include* sentences become more efficient.

	epoch	prec	recall	f1	f2	f3
mean	3.25	0.6494	0.9583	0.7711	0.8722	0.9129
var	2.33	0.0061	0.0003	0.0027	0.0005	0.0001

Table 3. Binary classification performances for 4-fold cross validation, where epoch is the epoch that generates the best-performing model.

report id	train bleu	train sari	epoch	bleu	sari
17	13.92	62.7	2	14.31	65.09
21	36.98	71.8	14	31.92	77.49
111	31.92	77.49	9	27.34	68.72
180	18.88	64.96	5	17.81	64.56

Table 4. Sentence simplification performances for the 4 test reports.

original_sentence	label	predicted
There is no evidence of filling defects in the pulmonary arteries to suspect pulmonary embolism.	There are no blood clots in the lungs.	No blood clots in the lungs.
There is stable trace pericardial effusion, likely physiologic.	There is small amount of fluid around the lungs.	There is fluid around the lungs.
There is redemonstration of upper lobes predominant centrilobular and paraseptal emphysema.	The lungs look like they are affected by emphysema.	There is emphysema in the lungs.
There are mild bibasilar subsegmental atelectasis.	The bottom of both lungs are not inflating well.	The lungs are not inflating well.
Scattered sub-4 mm lung nodules are stable.	Spots in the lungs are the same as before.	Spots in the lung look the same.
There is 4 mm right upper lobe pleural-based calcific density.	There is a hardened spot in the top of the right lung.	There is a small amount of fluid around the lungs.

Table 5. Example predictions for report 21.

Models & Results

For binary classification, we use the pre-trained model microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract to create an representation for each sentence. A linear layer with output dimension 2 is built on top of the representation. We divide the data into 4 folds with the same number of sentences, and record the precision, recall, and various f-beta scores for each epoch. f3 is used to select the best-performing model. Each fold is run with 3 random restarts and the average performances are reported in Table 3.

For sentence-level simplification, we use facebook/bart-base for conditional generation, where a language modeling head is added on top of the Bart model. In this task, we use SARI score to select the best-performing model. Four reports are selected as the test sets and the performances are reported in Table 4. Finally, example predictions for report 21 are shown in Table 5.

Future Work

Data: Currently, only ~15% of sentences from the 189 reports are annotated. The remaining data should be annotated and potentially extend to modalities beyond chest CTs.
Modeling: We only explore one model pre-trained on medical text in this project. Various other pre-trained models can also be utilized.
Evaluation: In this project, we find that SARI is insufficient in fully capturing the quality of simplified sentences, suggesting that improved automatic metrics and expert evaluations are necessary.

Conclusion

In conclusion, using state-of-the-art transformer models in radiology report summarization and simplification demonstrate promising results. Further effort in data annotation and model experimentation will be extremely valuable in pushing the system into real-life applications and improve patients' healthcare experiences.