Day 167: NLP Papers Summary – Ontology-Aware Clinical Abstractive Summarization

By Ryan 15th June 2020 No Comments

Objective and Contribution

Proposed an ontology-aware pointer-generator model for radiology report summarisation. Radiology report consists of two main sections: FINDINGS and IMPRESSION (summary). We use domain-specific ontology to improve content selection and this augmentation has shown to outperform SOTA results in terms of ROUGE scores. We also perform human evaluation on how good our generated summaries are in retaining salient information.

WHAT'S THE PROBLEM WITH EXISTING ABSTRACTIVE SUMMARISATION MODE ON CLINICAL NOTES?

They suffer from content selection and completeness. It is crucial for the summary of clinical notes to capture all the main diagnoses.

Methodology

We extend the original pointer-generator network with domain-specific ontology. Here, we link ontologies (UMLS or RadLex) to input clinical text to from a new encoded sequence. We do this by using a mapping function to output concepts if the input token appears in the ontology, otherwise we skip it. We also have a second biLSTM to encode the additional ontology terms and an additional context vector that uses the domain-ontology sequence. This additional context vector acts as additional global information to help the decoding process, which we have modified the decoder to accept the new ontological-aware context vector.

ONTOLOGIES

We use two ontologies: UMLS and RadLex. UMLS is a medical ontology that includes different procedures, conditions, symptoms, and many more. We specifically use QuickUMLS to extract UMLS concepts from the FINDINGS section. RadLex contains widely-used ontology of radiology terms and it's organised in a hierarchical structure. We use exact n-gram matching to identify important radiology entities.

Experiments and Results

Our evaluation dataset is 41066 radiology reports from MedStar Georgetown University Hospital. Our evaluation metric is ROUGE scores. Our baseline models are LSA, LexRank, vanilla pointer-generator model (with copy mechanism), and background-aware PG (encodes BACKGROUND section of the radiology report to improve summarisation).

RESULTS

Our ontology-aware models outperformed the other baseline models on both the dev and test set, with the RadLex ontology model achieving the highest performance. Both LexRank and LSA significantly underperformed, indicating that the most central sentences migh[illegible] not be important for the IMPRESSION summary. We also examine the attention weights t[illegible] ess if our ontology-aware decoder was able to attend more radiology terms. Figure 2 below showcase the attention weights of two samples. Relative to vanilla PG, our model was a to attend more radiological terms within the FINDINGS section, improving the summary generation process.

Table 1: ROUGE results on MedStar Georgetown University Hospital's development and test sets. Both the UMLS and RadLex ontology PG models are statistically better than the other models (paired t-test, $p < 0.05$).

	Development			Test		
Model	RG-1	RG-2	RG-L	RG-1	RG-2	RG-L
LexRank [5]	27.60	13.85	25.79	28.02	14.26	26.24
LSA [16]	28.04	14.68	26.15	28.16	14.71	26.27
PG [14]	36.60	21.73	35.40	37.17	22.36	35.45
Back. PG [19]	36.58	21.86	35.39	36.95	22.37	35.68
UMLS PG (ours)	37.41	22.23	36.10	37.98	23.14	36.67
RadLex PG (ours)	**37.64**	**22.45**	**36.33**	**38.42**	**23.29**	**37.02**

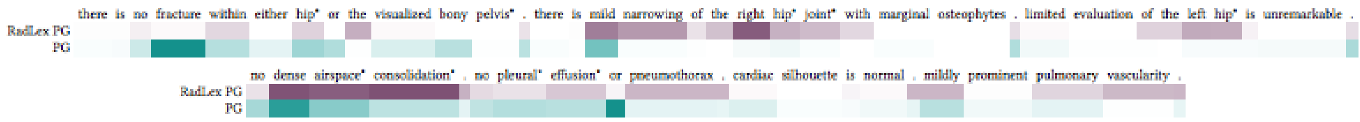

Figure 2: Average attention weight comparison between our approach (RadLex PG) and the baseline (PG). Color differences show to which term each model attends more while generating summary. RadLex concepts of depth 8 or lower are marked with *. Our approach attends to more RadLex terms throughout the document, allowing for more complete summaries.

HUMAN EVALUATION

We have a expert radiologist to manually evaluate 100 reports where each report has the radiology FINDINGS, the manually-written summary, PG-generated summary, and RadLex-ontology-regenerated summary. The radiologist score each IMPRESSION according to the following factors from 1 (worst) – 5 (best):

1. Readability
2. Accuracy
3. Completeness

The results are displayed below and show that our RadLex-generated summary improves completeness and accuracy while maintaining readability. There is a net gain of 10 reports that shown improvement in completeness between score 3 and 4. Our approach also has significantly less critical errors where only 5% of our generated summaries were scor nd 2. The radiologist also evaluated our RadLex-generated summaries independently to ide y how the summaries could improve further. Our RadLex-generated summaries was able to ect important points that are related to the RadLex terms and in some occasions, it was able to identify important points that the radiologist has missed. One problem is that our approach still generates repetitive sentences sometimes which we could mitigate by doing further pre-processing such as removing repetitive n-grams. Another problem is that our approach occasionally mix up the details leading to factual inconsistent, indicating that we could spend more time further developing our attention mechanism.

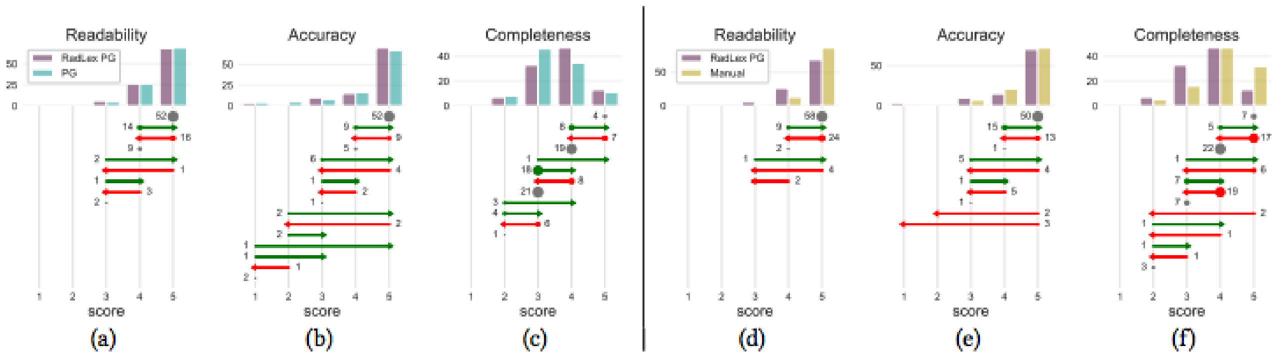

Figure 3: Histograms and arrow plots plot depicting differences between IMPRESSIONS of 100 manually-scored radiology reports. Although challenges remain to reach human parity for all metrics, our approach makes strong gains to address the problem of report completeness (c, f), as compared to the next leading summarization approach (PG).

Source: **https://arxiv.org/pdf/1905.05818.pdf**

Ryan

Data Scientist

Next Post

Day 168: Learn NLP With Me – Fast.Ai NLP Course

Previous Post

Day 166: NLP Papers Summary -

Publicly Available Clinical BERT Embeddings

– Understanding RNNs and Seq2Seq Translation

Data Science Natural Language Processing NLP Papers Summary

Day 365: NLP Papers Summary – A Survey on Knowledge Graph Embedding

Ryan
30th December 2020

Data Science Implementation Natural Language Processing

Day 364: Ryan's PhD Journey – OpenKE-PyTorch Library Analysis + code snippets for 11 KE models

Ryan
29th December 2020

Data Science Ryan's PhD Journey

Day 363: Ryan's PhD Journey – Literature Review – Knowledge Acquisition – 1st Passes XIV

Ryan
28th December 2020