Ontology-Aware Clinical Abstractive Summarization

Sean MacAvaney*
IRLab, Georgetown University
sean@ir.cs.georgetown.edu

Sajad Sotudeh*
IRLab, Georgetown University
sajad@ir.cs.georgetown.edu

Arman Cohan
Allen Institute for Artificial Intelligence
armanc@allenai.org

Nazli Goharian
IRLab, Georgetown University
nazli@ir.cs.georgetown.edu

Ish Talati
Department of Radiology,
Georgetown University
iat6@georgetown.edu

Ross W. Filice
MedStar Georgetown
University Hospital
ross.w.filice@medstar.net

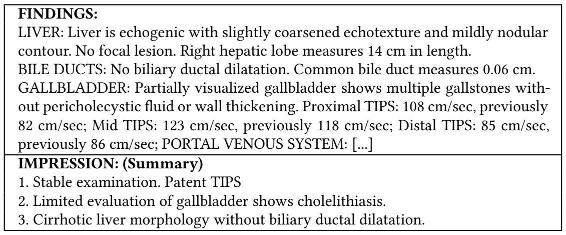
ABSTRACT

Automatically generating accurate summaries from clinical reports could save a clinician's time, improve summary coverage, and reduce errors. We propose a sequence-to-sequence abstractive summarization model augmented with domain-specific ontological information to enhance content selection and summary generation. We apply our method to a dataset of radiology reports and show that it significantly outperforms the current state-of-the-art on this task in terms of ROUGE scores. Extensive human evaluation conducted by a radiologist further indicates that this approach yields summaries that are less likely to omit important details, without sacrificing readability or accuracy.

ACM Reference Format:
Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish Talati, and Ross W. Filice. 2019. Ontology-Aware Clinical Abstractive Summarization. In *Proceedings of Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331319

1 INTRODUCTION

Clinical note summaries are critical to the clinical process. After writing a detailed note about a clinical encounter, practitioners often write a short summary called an IMPRESSION (example shown in Figure 1). This summary is important because it is often the primary document of the encounter considered when reviewing a patient's clinical history. The summary allows for a quick view of the most important information from the report. Automated summarization of clinical notes could save clinicians' time, and has the potential to capture important aspects of the note that the author might not have considered [7]. If high-quality summaries are generated frequently, the practitioner may only need to review the summary and occasionally make minor edits.



FINDINGS:
LIVER: Liver is echogenic with slightly coarsened echotexture and mildly nodular contour. No focal lesion. Right hepatic lobe measures 14 cm in length.
BILE DUCTS: No biliary ductal dilatation. Common bile duct measures 0.06 cm.
GALLBLADDER: Partially visualized gallbladder shows multiple gallstones without pericholecystic fluid or wall thickening. Proximal TIPS: 108 cm/sec, previously 82 cm/sec; Mid TIPS: 123 cm/sec, previously 118 cm/sec; Distal TIPS: 85 cm/sec, previously 86 cm/sec; PORTAL VENOUS SYSTEM: [...]

IMPRESSION: (Summary)
1. Stable examination. Patent TIPS
2. Limited evaluation of gallbladder shows cholelithiasis.
3. Cirrhotic liver morphology without biliary ductal dilatation.

Figure 1: Abbreviated example of radiology note and its summary.

Recently, neural abstractive summarization models have shown successful results [1, 11, 13, 14]. While promising in general domains, existing abstractive models can suffer from deficiencies in content accuracy and completeness [18], which is a critical issue in the medical domain. For instance, when summarizing a clinical note, it is crucial to include all the main diagnoses in the summary accurately. To overcome this challenge, we propose an extension to the pointer-generator model [14] that incorporates domain-specific knowledge for more accurate content selection. Specifically, we link entities in the clinical text with a domain-specific medical ontology (e.g., RadLex[1] or UMLS[2]), and encode them into a separate context vector, which is then used to aid the generation process. We train and evaluate our proposed model on a large collection of real-world radiology FINDINGS and IMPRESSIONS from a large urban hospital, MedStar Georgetown University Hospital. Results using the ROUGE evaluation metric indicate statistically significant improvements over existing state-of-the-art summarization models. Further extensive human evaluation by a radiology expert demonstrates that our method produces more complete summaries than the top-performing baseline, while not sacrificing readability or accuracy.

In summary, our contributions are: 1) An approach for incorporating domain-specific information into an abstractive summarization model, allowing for domain-informed decoding; and 2) Extensive automatic and human evaluation on a large collection of radiology notes, demonstrating the effectiveness of our model and providing insights into the qualities of our approach.

1.1 Related Work

Recent trends on abstractive summarization are based on sequence-to-sequence (seq2seq) neural networks with the incorporation of

*Both authors contributed equally to this research.

SIGIR '19, July 21–25, 2019, Paris, France
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331319

[1] RadLex version 3.10, http://www.radlex.org/Files/radlex3.10.xlsx
[2] https://www.nlm.nih.gov/research/umls/

attention [13], copying mechanism [14], reinforcement learning objective [8, 12], and tracking coverage [14]. While successful, a few recent studies have shown that neural abstractive summarization models can have high readability, but fall short in generating accurate and complete content [6, 18]. Content accuracy is especially crucial in medical domain. In contrast with prior work, we focus on improving summary completeness using a medical ontology. Gigioli et al. [8] used a reinforced loss for abstractive summarization in the medical domain, although their focus was headline generation from medical literature abstracts. Here, we focus on summarization of clinical notes where content accuracy and completeness are more critical. The most relevant work to ours is by Zhang et al. [19] where an additional section from the radiology report (BACKGROUND) is used to improve summarization. Extensive automated and human evaluation and analyses demonstrate the benefits of our proposed model in comparison with existing work.

2 MODEL

Pointer-generator network (PG). Standard neural approaches for abstractive summarization follow the seq2seq framework where an encoder network reads the input and a separate decoder network (often augmented with an attention mechanism) learns to generate the summary [17]. Bidirectional LSTMs (BiLSTMs) [9] are often used as the encoder and decoder. A more recent successful summarization model—called Pointer-generator network—allows the decoder to also directly copy text from the input in addition to generation [14]. Given a report $\mathbf{x} = \{x_1, x_2, ..., x_n\}$, the encoded input sequence $\mathbf{h} = \text{BiLSTM}(\mathbf{x})$, and the current decoding state $\mathbf{s_t} = \text{BiLSTM}(\mathbf{x}')[t]$, where $\mathbf{x}'$ is the input to the decoder (i.e., gold standard summary token at training or previously generated token at inference time), the model computes the attention weights over the input terms $\mathbf{a} = \text{softmax}(\mathbf{h}^\top \mathbf{W_1} \mathbf{s}^\top)$. The attention scores are employed to compute a context vector c which is a weighted sum over input $\mathbf{c} = \sum_i^n a_i \mathbf{h}_i$ that is used along with the output of the decoder BiLSTM to either generate the next term from a known vocabulary or copy the token from the input sequence with the highest attention value. We refer the reader to See et al. [14] for additional details on the pointer-generator architecture.

Ontology-aware pointer-generator (Ontology PG). In this work, we propose an extension of the pointer-generator network that allows us to leverage domain-specific knowledge encoded in an ontology to improve clinical summarization. We introduce a new encoded sequence $\mathbf{u} = \{u_1, ..., u_{n'}\}$ that is the result of linking an ontology $\mathscr{U}$ to the input texts. In other words, $\mathbf{u} = F_{\mathscr{U}}(\mathbf{x})$ where $F_{\mathscr{U}}$ is a mapping function, e.g., a simple mapping function that only outputs a word sequence if it appears in the ontology and otherwise skips it. We then use a second BiLSTM to encode this additional ontology terms similar to the way the original input is encoded $\mathbf{h}_u = BiLSTM(\mathbf{u})$. We then calculate an additional context vector $\mathbf{c}'$ which includes the domain-ontology information:

$$\mathbf{a}' = \text{softmax}(\mathbf{h}_\mathbf{u}^\top \mathbf{W_2} \mathbf{s}^\top); \quad \mathbf{c}' = \sum\nolimits_i^{n'} a'_i \mathbf{u}_i \tag{1}$$

The second context vector acts as additional global information to aid the decoding process, and is akin to how Zhang et al. [19] include BACKGROUND information from the report. We modify the decoder BiLSTM to include the ontology-aware context vector in the decoding process. Recall that an LSTM network controls the flow of its previous state and the current input using several gates (input gate $\mathbf{i}$, forget gate $\mathbf{f}$, and output gate $\mathbf{o}$), where each of these gates are vectors calculated according to an additive combination of the previous LSTM state and current input. For example, for the forget gate we have: $\mathbf{f}_t = \tanh(W_f[s_{t-1}; x'_t] + b)$ where s_{t-1} is the previous decoder state and x'_t is the decoder input, and ";" shows concatenation (for more details on LSTMs refer to [9]). The ontology-aware context vector c' is passed as additional input to this function for all the LSTM gates: e.g., for the forget gate we will have: $\mathbf{f}_t = \tanh(W_f[s_{t-1}; x'_t; c'] + b)$. This intuitively guides the information flow in the decoder using the ontology information.

3 EXPERIMENTAL SETUP

We train and evaluate our model on a dataset of 41,066 real-world radiology reports from MedStar Georgetown University Hospital containing radiology reports with a variety of imaging modalities (e.g., x-rays, CT scans, etc). The dataset is randomly split into 80-10-10 train-dev-test splits. Each report describes clinical FINDINGS about a specific diagnostic case, and an IMPRESSION summary (as shown in Figure 1). The FINDINGS sections are 136.6 tokens on average and the IMPRESSION sections are 37.1 tokens on average. Performing cross-institutional evaluation is challenging and beyond the scope of this work due to the varying nature of reports between institutions. For instance, the public Indiana University radiology dataset [4] consists only of chest x-rays, and has much shorter reports (average length of FINDINGS: 40.0 tokens; average length of IMPRESSIONS: 10.5 tokens). Thus, in this work, we focus on summarization within a single institution.

Ontologies. We employ two ontologies in this work. UMLS is a general medical ontology maintained by the US National Library of Medicine and includes various procedures, conditions, symptoms, body parts, etc. We use QuickUMLS [15] (a fuzzy UMLS concept matcher) with a Jaccard similarity threshold of 0.7 and a window size of 3 to extract UMLS concepts from the radiology FINDINGS. We also evaluate using an ontology focused on radiology, RadLex, which is a widely-used ontology of radiological terms maintained by the Radiological Society of North America. It consists of 68,534 radiological concepts organized according to a hierarchical structure. We use exact n-gram matching to find important radiological entities, only considering RadLex concepts at a depth of 8 or greater.[3] In pilot studies, we found that the entities between depths 8 and 20 tend to represent concrete entities (e.g., 'thoracolumbar spine region') rather than abstract categories (e.g., 'anatomical entity').

Comparison. We compare our model to well-established extractive baselines as well as the state-of-the-art abstractive summarization models.

- **LSA** [16]: An extractive vector-space summarization model based on Singular Value Decomposition (SVD).
- **LexRank** [5]: An extractive method which employs graph-based centrality ranking of the sentence.[4]
- **Pointer-Generator (PG)** [14]: An abstractive seq2seq attention summarization model that incorporates a copy mechanism to directly copy text from input where appropriate.

[3] The maximum tree depth is 20.

[4] For LSA and LexRank, we use the Sumy implementation (https://pypi.python.org/pypi/sumy) with the top 3 sentences.

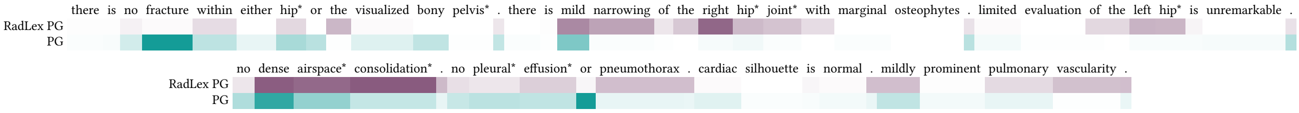


Figure 2: Average attention weight comparison between our approach (RadLex PG) and the baseline (PG). Color differences show to which term each model attends more while generating summary. RadLex concepts of depth 8 or lower are marked with *. Our approach attends to more RadLex terms throughout the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

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**

- **Background-Aware Pointer-Generator (Back. PG)** [19]: An extension of PG, which is specifically designed to improve radiology note summarization by encoding the Background section of the report to aid the decoding process.[5]

Parameters and training. We use 100-dimensional GloVe embeddings pre-trained over a large corpus of 4.5 million radiology reports [19], a 2-layer BiLSTM encoder with a hidden size of 100, and a 1-layer LSTM decoder with the hidden size of 200. At inference time, we use beam search with beam size of 5. We use a dropout of 0.5 in all models, and train to optimize negative log-likelihood loss using the Adam optimizer [10] and a learning rate of 0.001.

4 RESULTS AND ANALYSIS

4.1 Experimental results

Table 1 presents ROUGE evaluation results of our model compared with the baselines (as compared to human-written IMPRESSIONS). The extractive summarization methods (LexRank and LSA) perform particularly poorly. This may be due to the fact that these approaches are limited to simply selecting sentences from the text, and that the most central sentences may not be the most important for building an effective IMPRESSION summary. Interestingly, the Back. PG approach (which uses the BACKGROUND section of the report to guide the decoding process) is ineffective on our dataset. This may be due to differences in conventions across institutions, such as what information is included in a report’s BACKGROUND and what is considered important to include in its IMPRESSION.

We observe that our Ontology-Aware models (UMLS PG and RadLex PG) significantly outperform all other approaches (paired t-test, $p < 0.05$) on both the development and test sets. The RadLex model slightly outperforms the UMLS model, suggesting that the radiology-specific ontology is beneficial (though the difference between UMLS and RadLex is not statistically significant). We also experimented incorporating both ontologies in the model simultaneously, but it resulted in slightly lower performance (1.26% lower than the best model on ROUGE-1). To verify that including ontological concepts in the decoder helps the model identify and focus on more radiology terms, we examined the attention weights. In Figure 2, we show attention plots for two reports, comparing the attention of our approach and PG. The plots show that our approach results in attention weights being shared across radiological terms throughout the FINDINGS, potentially helping the model to capture a more complete summary.

4.2 Expert human evaluation

While our approach surpasses state-of-the-art results on our dataset in terms of ROUGE scores, we recognize the limitations of the ROUGE framework for evaluating summarization [2, 3]. To gain better insights into how and why our methodology performs better, we also conduct expert human evaluation. We had a domain expert (radiologist) who is familiar with the process of writing radiological FINDINGS and IMPRESSIONS evaluate 100 reports. Each report consists of the radiology FINDINGS, one manually-written IMPRESSION, one IMPRESSION generated using PG, and one IMPRESSION generated using our ontology PG method (with RadLex). In each sample, the order of the IMPRESSIONS are shuffled to avoid bias between samples. Samples were randomly chosen from the test set, one from each of 100 evenly-spaced bins sorted by our system’s ROUGE-1 score. The radiologist was asked to score each IMPRESSION in terms of the following on a scale of 1 (worst) to 5 (best):

- **Readability.** Impression is understandable (5) or gibberish (1).
- **Accuracy.** Impression is fully accurate (5), or contains critical errors (1).
- **Completeness.** Impression contains all important information (5), or is missing important points (1).

We present our manual evaluation results using histograms and arrow plots in Figure 3. The histograms indicate the score distributions of each approach, and the arrows indicate how the scores changed. The starting points of an arrow indicates the score of an IMPRESSION we compare to (either the human-written, or the summary generated by PG). The head of an arrow indicates the score of our approach. The numbers next to each arrow indicate how many reports made the transition. The figure shows that our approach improves completeness considerably, while maintaining the readability and accuracy. The major improvement in completeness is between the score of 3 and 4, where there is a net gain of 10 reports. Completeness is particularly important because it is where

[5] Using the author’s code at github.com/yuhaozhang/summarize-radiology-findings

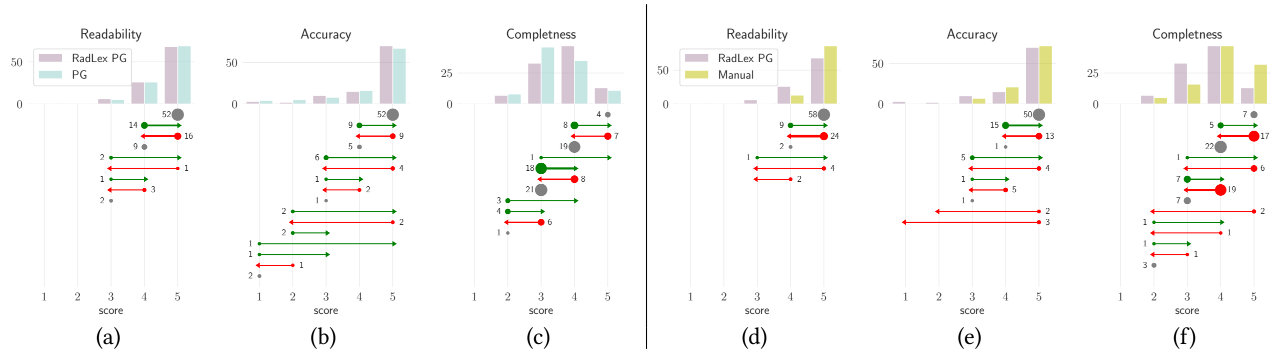


Figure 3: Histograms and arrow plots plot depicting differences between IMPRESSIONS of 100 manually-scored radiology reports. Although challenges remain to reach human parity for all metrics, our approach makes strong gains to address the problem of report completeness (c, f), as compared to the next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

existing summarization models—such as PG—are currently lacking, as compared to human performance. Despite the remaining gap between human and generated completeness, our approach yields considerable gains toward human-level completeness. Our model is nearly as accurate as human-written summaries, only making critical errors (scores of 1 or 2) in 5% of the cases evaluated, as compared to 8% of cases for PG. No critical errors were found in the human-written summaries, although the human-written summaries go through a manual review process to ensure accuracy.

The expert annotator furthermore conducted blind qualitative analysis to gain a better understanding of when our model is doing better and how it can be further improved. In line with the completeness score improvements, the annotator noted that in many cases our approach is able to identify pertinent points associated with RadLex terms that were missed by the PG model. In some cases, such as when the author picked only one main point, our approach was able to pick up important items that the author missed. Interestingly, it also was able to include specific measurement details better than the PG network, even though these measurements do not appear in RadLex. Although readability is generally strong, our approach sometimes generates repetitive sentences and syntactical errors more often than humans. These could be addressed in future work with additional post-processing heuristics such as removing repetitive n-grams as done in [12]. In terms of accuracy, our approach sometimes mixes up the "left" and "right" sides. This often occurs with FINDINGS that have mentions of both sides of a specific body part. Multi-level attention (e.g., [1]) could address this by forcing the model to focus on important segments of the text. There were also some cases where our model under-performed in terms of accuracy and completeness due to synonymy that is not captured by RadLex. For instance, in one case our model did identify torsion, likely due to the fact that in the FINDINGS section it was referred to as *twisting* (a term that does not appear in RadLex).

5 CONCLUSION

In this work, we present an approach for informing clinical summarization models of ontological information. This is accomplished by providing an encoding of ontological terms matched in the original text as an additional feature to guide the decoding. We find that our system exceeds state-of-the-art performance at this task, producing summaries that are more comprehensive than those generated by other methods, while not sacrificing readability or accuracy.

ACKNOWLEDGMENTS

We thank the additional residents, Lee McDaniel and Marlie Philiossaint, for their contributive evaluations, as well as anonymous reviewers for their useful feedback.

REFERENCES

[1] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. In *NAACL-HLT*.

[2] Arman Cohan and Nazli Goharian. 2016. Revisiting Summarization Evaluation for Scientific Articles. *Proc. of 11th Conference on LREC* (2016), 806–813.

[3] John M. Conroy and Hoa Trang Dang. 2008. Mind the Gap: Dangers of Divorcing Evaluations of Summary Content from Linguistic Quality. In *COLING*.

[4] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association* (2015).

[5] Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based Lexical Centrality As Salience in Text Summarization. *J. Artif. Int. Res.* (2004), 457–479.

[6] Sebastian Gehrmann, Yuntian Deng, and Alexander M. Rush. 2018. Bottom-Up Abstractive Summarization. In *EMNLP*.

[7] Esteban F Gershanik, Ronilda Lacson, and Ramin Khorasani. 2011. Critical finding capture in the impression section of radiology reports. In *AMIA*.

[8] Paul Gigioli, Nikhita Sagar, Anand S. Rao, and Joseph Voyles. 2018. Domain-Aware Abstractive Text Summarization for Medical Documents. In *IEEE BIBM*.

[9] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[10] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

[11] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *CoNLL*.

[12] Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A Deep Reinforced Model for Abstractive Summarization. *CoRR* (2017).

[13] Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In *EMNLP*.

[14] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. *ACL* (2017).

[15] Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR Workshop, SIGIR*.

[16] Josef Steinberger and Karel Ježek. 2004. Using latent semantic analysis in text summarization and summary evaluation. In *ISIM*.

[17] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*. 3104–3112.

[18] Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2017. Challenges in Data-to-Document Generation. In *EMNLP*.

[19] Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018. Learning to Summarize Radiology Findings. In *EMNLP Workshop on Health Text Mining and Information Analysis*.