Attend to Medical Ontologies: Content Selection for Clinical Abstractive Summarization

Sajad Sotudeh[1], **Nazli Goharian**[1], and **Ross W. Filice**[2]

[1]IR Lab, Georgetown University, Washington DC 20057, USA
{sajad, nazli}@ir.cs.georgetown.edu
[2]MedStar Georgetown University Hospital, Washington DC 20007, USA
ross.w.filice@medstar.net

Abstract

Sequence-to-sequence (seq2seq) network is a well-established model for text summarization task. It can learn to produce readable content; however, it falls short in effectively identifying key regions of the source. In this paper, we approach the content selection problem for clinical abstractive summarization by augmenting salient ontological terms into the summarizer. Our experiments on two publicly available clinical data sets (107,372 reports of MIMIC-CXR, and 3,366 reports of OpenI) show that our model statistically significantly boosts state-of-the-art results in terms of ROUGE metrics (with improvements: 2.9% RG-1, 2.5% RG-2, 1.9% RG-L), in the healthcare domain where any range of improvement impacts patients' welfare.

1 Introduction

Radiology reports convey the detailed observations along with the significant findings about a medical encounter. Each radiology report contains two important sections:[1] FINDINGS that encompasses radiologist's detailed observations and interpretation of imaging study, and IMPRESSION summarizing the most critical findings. IMPRESSION (usually couple of lines and thrice smaller than finding) is considered as the most integral part of report (Ware et al., 2017) as it plays a key role in communicating critical findings to referring clinicians. Previous studies have reported that clinicians mostly read the IMPRESSION as they have less time to review findings, particularly those that are lengthy or intricate (Flanders and Lakhani, 2012; Xie et al., 2019).

In clinical setting, generating IMPRESSION from FINDINGS can be subject to errors (Gershanik et al., 2011; Brady, 2016). This fact is especially crucial when it comes to healthcare domain where even the smallest improvement in generating IMPRESSION can improve patients' well-being. Automating the process of impression generation in radiology reporting would save clinicians' read time and decrease fatigue (Flanders and Lakhani, 2012; Kovacs et al., 2018) as clinicians would only need to proofread summaries or make minor edits.

Previously, MacAvaney et al. (2019) showed that augmenting the summarizer with entire ontology (i.e., clinical) terms within the FINDINGS can improve the content selection and summary generation to some noticeable extent. Our findings, further, suggest that radiologists select *significant* ontology terms, but not all such terms, to write the IMPRESSION. Following this paradigm, we hypothesize that selecting the most *significant* clinical terms occurring in the FINDINGS and then incorporating them into the summarization would improve the final IMPRESSION generation. We further examine if refining FINDINGS word representations according to the identified clinical terms would result in improved IMPRESSION generation.

Overall, the contributions of this work are twofold: (i) We propose a novel seq2seq-based model to incorporate the salient clinical terms into the summarizer (§3.2). We pose copying likelihood of a word as an indicator of its saliency in terms of forming IMPRESSION, which can be learned via a sequence-tagger (§3.1); (ii) Our model statistically significantly improves over the competitive baselines on MIMIC-CXR publicly available clinical dataset. To evaluate the cross-organizational transferability, we further evaluate our model on another publicly available clinical dataset (OpenI) (§5).

2 Related Work

Few prior studies have pointed out that although seq2seq models can effectively produce readable content, they perform poorly at selecting salient

[1]Depending on institution, radiology reports may or may not include other fields such as BACKGROUND.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1899–1905

content to include in the summary (Gehrmann et al., 2018; Lebanoff et al., 2019). Many attempts have been made to tackle this problem (Zhou et al., 2017; Lin et al., 2018; Hsu et al., 2018; Lebanoff et al., 2018; You et al., 2019). For example, Zhou et al. (2017) used sentence representations to filter secondary information of word representation. Our work is different in that we utilize ontology representations produced by an additional encoder to filter word representations. Gehrmann et al. (2018) utilized a data-efficient content selector, by aligning source and target, to restrict the model's attention to likely-to-copy phrases. In contrast, we use the content selector to find domain knowledge alignment between source and target. Moreover, we do not focus on model attention here, but on rectifying word representations.

Extracting clinical findings from clinical reports has been explored previously (Hassanpour and Langlotz, 2016; Nandhakumar et al., 2017). For summarizing radiology reports, Zhang et al. (2018) recently used a separate RNN to encode a section of radiology report.[2] Subsequently, MacAvaney et al. (2019) extracted clinical ontologies within the FINDINGS to help the model learn these useful signals by guiding decoder in generation process. Our work differs in that we hypothesize that all of the ontological terms in the FINDINGS are not equally important, but there is a notion of *odds of saliency* for each of these terms; thus, we focus on refining the FINDINGS representations.

3 Model

Our model consists of two main components: (1) a content selector to identify the most salient ontological concepts specific to a given report, and (2) a summarization model that incorporates the identified ontology terms within the FINDINGS into the summarizer. The summarizer refines the FINDINGS word representation based on salient ontology word representation encoded by a separate encoder.

3.1 Content Selector

The content selection problem can be framed as a word-level extraction task in which the aim is to identify the words within the FINDINGS that are likely to be copied into the IMPRESSION. We tackle this problem through a sequence-labeling approach. We align FINDINGS and IMPRESSION to obtain required data for sequence-labeling task. To this end, let $b_1, b_2, ..., b_n$ be the binary tags over the FINDINGS terms $\mathbf{x} = \{x_1, x_2, ..., x_n\}$, with n being the length of the FINDINGS. We tag word x_i with 1 if it meets two criteria simultaneously: (1) it is an ontology term, (2) it is directly copied into IMPRESSION, and 0 otherwise. At inference, we characterize the copying likelihood of each FINDINGS term as a measure of its saliency.

Recent studies have shown that contextualized word embeddings can improve the sequence-labeling performance (Devlin et al., 2019; Peters et al., 2018). To utilize this improvement for the content selection, we train a bi-LSTM network on top of the BERT embeddings with a softmax activation function. The content selector is trained to maximize log-likelihood loss with the maximum likelihood estimation. At inference, the content selector calculates the selection probability of each token in the input sequence. Formally, let $\mathcal{O}$ be the set of ontological words which the content selector predicts to be copied into the IMPRESSION:

$$\mathcal{O} = \{o_i | o_i \in F_{\mathcal{U}}(\mathbf{x}) \wedge p_{o_i} \geq \epsilon\} \quad (1)$$

where $F_{\mathcal{U}}(\mathbf{x})$ is a mapping function that takes in FINDINGS tokens and outputs word sequences from input tokens if they appear in the ontology (i.e., RadLex) [3], and otherwise skips them. p_{o_i} denotes the selection probability of ontology word o_i, and $\epsilon \in [0, 1]$ is the copying threshold.

3.2 Summarization Model

3.2.1 Encoders

We exploit two separate encoders: (1) findings encoder that takes in the FINDINGS, and (2) ontology encoder that maps significant ontological terms identified by the content selector to a fix vector known as ontology vector. The findings encoder is fed with the embeddings of FINDINGS words, and generates word representations $\mathbf{h}$. Then, a separate encoder, called ontology encoder, is used to process the ontology terms identified by the content selector and produce associated representations $\mathbf{h^o}$.

$$\begin{aligned} \mathbf{h} &= \text{Bi-LSTM}(\mathbf{x}) \\ \mathbf{h^o} &= \text{LSTM}(\mathcal{O}) \end{aligned} \quad (2)$$

where $\mathbf{x}$ is the FINDINGS text, $\mathcal{O}$ is the set of ontology terms occurring in the FINDINGS and identified by the content selector, $\mathbf{h^o} = \{h^o_1, h^o_2, ..., h^o_l\}$ is the

[2] BACKGROUND field.

[3] RadLex version 3.10, http://www.radlex.org/Files/radlex3.10.xlsx

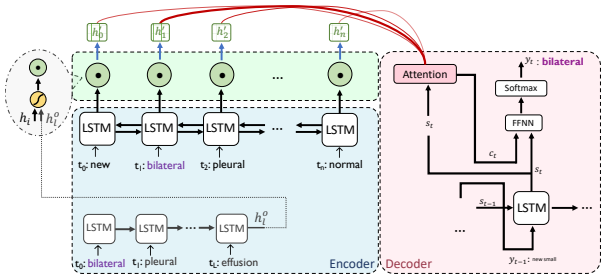

Figure 1: Overview of our summarization model. As shown, "bilateral" in the FINDINGS is a significant ontological term which has been encoded into the ontology vector. After refining FINDINGS word representation, the decoder computes attention weight (highest on "bilateral") and generates it in the IMPRESSION.

word representations yielded from the ontology encoder. Note that h_l^o–called ontology vector– is the last hidden state containing summarized information of significant ontologies in the FINDINGS.

3.2.2 Ontological Information Filtering

Although de facto seq2seq frameworks implicitly model the information flow from encoder to decoder, the model should benefit from explicitly modeling the selection process. To this end, we implement a filtering gate on top of the findings encoder to refine the FINDINGS word representations according to the significant ontology terms within the FINDINGS and produce ontology-aware word representations. Specifically, the filtering gate receives two vectors: the word hidden representation h_i that has the contextual information of word x_i, and the ontology vector h_l^o including the overal information of significant ontology words within the FINDINGS. The filtering gate processes these two vectors through a liner layer with Sigmoid activation function. We then compute the ontology-aware word hidden representation h'_i, given the source word hidden representation h_i and the associated filtering gate F_i.

$$\begin{aligned} F_i &= \sigma(W_h[h_i; h_l^o] + b) \\ h'_i &= h_i \odot F_i \end{aligned} \tag{3}$$

where W_h is the weight matrix, b denotes the bias term, and $\odot$ denotes element-wise multiplication.

3.2.3 Impression Decoder

We use an LSTM network as our decoder to generate the IMPRESSION iteratively. In this sense, the decoder computes the current decoding state $\mathbf{s_t} = \mathbf{LSTM}(\mathbf{s_{t-1}}, \mathbf{y_{t-1}})$, where $\mathbf{y_{t-1}}$ is the input to the decoder (human-written summary tokens at training, or previously generated tokens at inference) and $\mathbf{s_{t-1}}$ is the previous decoder state. The decoder also computes an attention distribution $\mathbf{a} = \mathrm{Softmax}(\mathbf{h'}^\top \mathbf{V} \mathbf{s}^\top)$ with $\mathbf{h'}$ being the ontology-aware word representations. The attention weights are then used to compute the context vector $\mathbf{c_t} = \sum_i^n a_i \mathbf{h'_i}$ where n is the length of the FINDINGS. Finally, the context vector and decoder output are used to either generate the next token from the vocabulary or copy it from the FINDINGS.

4 Experiments

4.1 Dataset and Ontologies

MIMIC-CXR. This collection (Johnson et al., 2019) is a large publicly available dataset of radiology reports. Following similar report preprocessing as done in (Zhang et al., 2018), we obtained 107,372 radiology reports. For tokenization, we used ScispaCy (Neumann et al., 2019). We randomly split the dataset into 80%(85,898)-10%(10,737)-10%(10,737) train-dev-test splits.

OpenI. A public dataset from the Indiana Network for Patient Care (Demner-Fushman et al., 2016) with 3,366 reports. Due to small size, it is not suitable for training; we use it to evaluate the cross-organizational transferability of our model and baselines.

Ontologies. We use RadLex, a comprehensive radiology lexicon, developed by Radiological Society of North America (RSNA), including 68,534 radiological terms organized in hierarchical structure.

4.2 Baselines

We compare our model against both known and state-of-the-art extractive and abstractive models.

- **LSA** (Steinberger and Ježek, 2004): An extractive vector-based model that employs Sigular Value Decomposition (SVD) concept.

- **NeuSum** (Zhou et al., 2018): A state-of-the-art extractive model that integrates the process of source sentence scoring and selection.[4]

- **Pointer-Generator (PG)** (See et al., 2017): An abstractive summarizer that extends ses2seq networks by adding a copy mechanism that allows for directly copying tokens from the source.

- **Ontology-Aware Pointer-Generator (Ont. PG)** (MacAvaney et al., 2019): An extension of

[4]We use open code at https://github.com/magic282/NeuSum with default hyper-parameters.

Method	RG-1	RG-2	RG-L
LSA	22.21	11.17	20.80
NEUSUM	23.97	12.82	22.61
PG	51.20	39.13	50.16
Ont. PG	51.84	39.59	50.72
BUS	52.04	39.69	50.83
Ours (this work)	**53.57***	**40.78***	**51.81***

Table 1: ROUGE results on MIMIC-CXR. * shows the statistical significance (paired t-test, $p < 0.05$).

PG model that first encodes *entire* ontological concepts within FINDINGS, then uses the encoded vector to guide decoder in summary decoding process.

- **Bottom-Up Summarization (BUS)** (Gehrmann et al., 2018): An abstractive model which makes use of a content selector to constrain the model's attention over source terms that have a good chance of being copied into the target.[5]

4.3 Parameters and Training

We use SCIBERT model (Beltagy et al., 2019) which is pre-trained over biomedical text. We employ 2-layer bi-LSTM encoder with hidden size of 256 upon BERT model. The dropout is set to 0.2. We train the network to minimize cross entropy loss function, and optimize using Adam optimizer (Kingma and Ba, 2015) with learning rate of $2e^{-5}$.

For the summarization model, we extended on the open base code by Zhang et al. (2018) for implementation.[6] We use 2-layer bi-LSTM, 1-layer LSTM as findings encoder, ontology encoder, and decoder with hidden sizes of 200 and 100, respectively. We also exploit 100d GloVe embeddings pretrained on a large collection of 4.5 million radiology reports (Zhang et al., 2018). We train the network to optimize negative log likelihood with Adam optimizer and a learning rate of 0.001.

5 Results and Discussion

5.1 Experimental Results

Table. 1 shows the ROUGE scores of our model and baseline models on MIMIC-CXR, with human-written IMPRESSIONS as the ground truth. Our model significantly outperforms all the baselines on all ROUGE metrics with 2.9%, 2.5%, and 1.9% improvements for RG-1, RG-2, and RG-L, respectively. While NEUSUM outperforms the non-neural LSA in extractive setting, the extractive models lag behind the abstractive methods considerably, suggesting that human-written impressions are formed by abstractively selecting information from the findings, not merely extracting source sentences. When comparing Ont. PG with our model, it turns out that indeed our hypothesis is valid that a pre-step of identifying significant ontological terms can improve the summary generation substantially. As pointed out earlier, we define the saliency of an ontological term by its copying probability.

As expected, BUS approach achieves the best results among the baseline models by constraining decoder's attention over odds-on-copied terms, but still underperforms our model. This may suggest that the intermediate stage of refining word representations based on the ontological word would lead to a better performance than superficially restricting attention over the salient terms. Table. 3 shows the effect of content selector on the summarization model. For the setting without content selector, we encode all ontologies within the FINDINGS. As shown, our model statistically significantly improves the results on RG-1 and RG-2.

To further evaluate the transferability of our model across organizations, we perform an evaluation on OpenI with our best trained model on MIMIC-CXR. As shown in Table. 2, our model significantly outperforms the top-performing abstractive baseline model suggesting the promising cross-organizational transferability of our model.

Method	RG-1	RG-2	RG-L
BUS	40.02	21.89	39.37
Ours (this work)	**40.88***	**24.44***	**40.37***

Table 2: ROUGE results on Open-I dataset, comparing our model with the best-performing baseline. * shows the statistical significance (paired t-test, $p < 0.05$).

Setting	RG-1	RG-2	RG-L
w/o Cont. Sel.	52.47	40.11	51.39
w/ Cont. Sel.	**53.57***	**40.78***	**51.81**

Table 3: ROUGE results showing the impact of content selector in summarization model. * shows the statistical significance (paired t-test, $p < 0.05$).

[5] We re-implemented the BUS model.

[6] https://github.com/yuhaozhang/summarize-radiology-findings

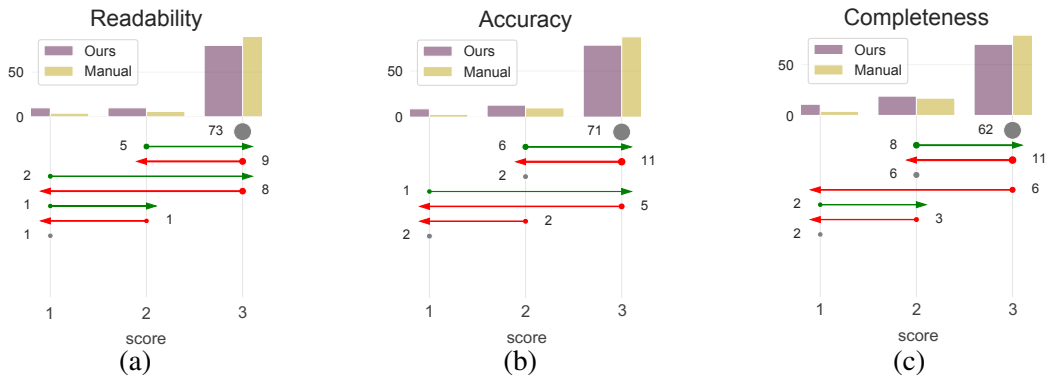

Figure 2: Histograms and arrow plots showing differences between IMPRESSION of 100 manually-scored radiology reports. Although challenges remain to reach human parity for all metrics, 81% (a), 82% (b), and 80% (c) of our system-generated Impressions are as good as human-written Impressions across different metrics.

5.2 Expert Evaluation

While our approach achieves the best ROUGE scores, we recognize the limitation of this metric for summarization task (Cohan and Goharian, 2016). To gain a better understanding of qualities of our model, we conducted an expert human evaluation. To this end, we randomly sampled 100 system-generated Impressions with their associated gold from 100 evenly-spaced bins (sorted by our system's RG-1) of MIMIC-CXR dataset. The Impressions were shuffled to prevent potential bias. We then asked three experts [7] to score the given Impressions independently on a scale of 1-3 (worst to best) for three metrics: *Readability.* understandable or nonsense; *Accuracy.* fully accurate, or containing critical errors; *Completeness.* having all major information, or missing key points.

Figure. 2 presents the human evaluation results using histograms and arrow plots as done in (MacAvaney et al., 2019), comparing our system's Impressions versus human-written Impressions. The histograms indicate the distribution of scores, and arrows show how the scores changed between ours and human-written. The tail of each arrow shows the score of human-written IMPRESSION , and its head indicates the score of our system's IMPRESSION. The numbers next to the tails express the count of Impressions that gained score of $\mathbf{s}'$ by ours and $\mathbf{s}$ by gold. [8] We observed that while there is still a gap between the system-generated and human-written Impressions, *over 80%* of our system-generated Impressions are as good [9] as the associated human-written Impressions. Specifically, 73% (readability), and 71% (accuracy) of our system-generated Impressions ties with human-written Impressions, both achieving full-score of 3; nonetheless, this percentage is 62% for completeness metric. The most likely explanation of this gap is that deciding which findings are more important (i.e., should be written into Impression) is either subjective, or highly correlates with the institutional training purposes. Hence, we recognize cross-organizational evaluations in terms of Impression completeness as a challenging task. We also evaluated the inter-rater agreement using Fleiss' Kappa (Fleiss, 1971) for our system's scores and obtained 52% for readability, 47% for accuracy, and 50% for completeness, all of which are characterized as moderate agreement rate.

[7] Two radiologists and one medical student.

[8] $\mathbf{s}, \mathbf{s}' \in \{1, 2, 3\}$

[9] Either tied or improved.

6 Conclusion

We proposed an approach to content selection for abstractive text summarization in clinical notes. We introduced our novel approach to augment standard summarization model with significant ontological terms within the source. Content selection problem is framed as a word-level sequence-tagging task. The intrinsic evaluations on two publicly available real-life clinical datasets show the efficacy of our model in terms of ROUGE metrics. Furthermore, the extrinsic evaluation by domain experts further reveals the qualities of our system-generated summaries in comparison with gold summaries.

Acknowledgement

We thank Arman Cohan for his valuable comments on this work. We also thank additional domain expert evaluators: Phillip Hyuntae Kim, and Ish Talati.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *EMNLP*.

Adrian P. Brady. 2016. Error and discrepancy in radiology: inevitable or avoidable? In *Insights into Imaging*.

Arman Cohan and Nazli Goharian. 2016. Revisiting summarization evaluation for scientific articles. *Proc. of 11th Conference on LREC*, pages 806–813.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer K. Antani, George R. Thoma, and Clement J. McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association : JAMIA*, 23 2:304–10.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Adam E. Flanders and Paras Lakhani. 2012. Radiology reporting and communications: a look forward. *Neuroimaging clinics of North America*, 22 3:477–96.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters.

Sebastian Gehrmann, Yuntian Deng, and Alexander M. Rush. 2018. Bottom-up abstractive summarization. In *EMNLP*.

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

Saeed Hassanpour and Curtis P. Langlotz. 2016. Information extraction from multi-institutional radiology reports. *Artificial intelligence in medicine*, 66:29–39.

Wan Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. In *ACL*.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. 2019. Mimic-cxr: A large publicly available database of labeled chest radiographs. *ArXiv*, abs/1901.07042.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Mark D. Kovacs, Maximilian Y Cho, Philip F. Burchett, and Michael A. Trambert. 2018. Benefits of integrated ris/pacs/reporting due to automatic population of templated reports. *Current problems in diagnostic radiology*, 48 1:37–39.

Logan Lebanoff, Kaiqiang Song, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019. Scoring sentence singletons and pairs for abstractive summarization. In *ACL*.

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the neural encoder-decoder framework from single to multi-document summarization. In *EMNLP*.

Junyang Lin, Xu Sun, Shuming Ma, and Qi Su. 2018. Global encoding for abstractive summarization. In *ACL*.

Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish Talati, and Ross W. Filice. 2019. Ontology-aware clinical abstractive summarization. *SIGIR*.

Nidhin Nandhakumar, Ehsan Sherkat, Evangelos E. Milios, Hong Gu, and Michael Butler. 2017. Clinically significant information extraction from radiology reports. In *DocEng*.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. In *BioNLP@ACL*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*.

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

Jeffrey B Ware, Saurabh W. Jha, Jenny K Hoang, Stephen R Baker, and Jill Wruble. 2017. Effective radiology reporting. *Journal of the American College of Radiology : JACR*, 14 6:838–839.

Zhe Xie, Yuanyuan Yang, Mingqing Wang, Ming Hui Li, Haozhe Huang, Dezhong Zheng, Rong Shu, and Tonghui Ling. 2019. Introducing information extraction to radiology information systems to improve the efficiency on reading reports. *Methods of information in medicine*, 58 2-03:94–106.

Yongjian You, Weijia Jia, Tianyi Liu, and Wenmian Yang. 2019. Improving abstractive document summarization with salient information modeling. In *ACL*.

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*.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *ACL*.

Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. 2017. Selective encoding for abstractive sentence summarization. In *ACL*.