Coherent and Concise Radiology Report Generation via Context Specific Image Representations and Orthogonal Sentence States

Litton J Kurisinkel, Ai Ti Aw, Nancy F. Chen
Institute for Infocomm Research, A*STAR, Singapore
litton_kurisinkel, aaiti, nfychen@i2r.a-star.edu.sg

Abstract

Neural models for text generation are often designed in an end-to-end fashion, typically with zero control over intermediate computations, limiting their practical usability in downstream applications. In this work, we incorporate explicit means into neural models to ensure topical continuity, content comprehensiveness and informativeness of automatically generated radiology reports. We propose a method to compute image representations specific to each sentential context to minimize hallucination caused by sequence-to-sequence approaches and to further eliminate redundant content by exploiting diverse sentence states. We conduct experiments to generate radiology reports from medical images of chest x-rays using MIMIC-CXR. Our model outperforms baselines by up to 18% and 29% respective in the evaluation for informativeness and content ordering respectively on objective metrics and 16% on human validations.

1 Introduction

Presenting information in text format has been critical to the development of human civilizations. Thus text generation is an important field in artificial intelligence and natural language processing, where the input to such natural language generation models could take on the form of text, graphs, images or database records (Koncel-Kedziorski et al., 2019; See et al., 2017; Kinghorn et al., 2018).

Recent advancements in natural language generation has been propelled by end-to-end neural models (e.g. (Chopra et al., 2016)), which has strong capabilities to learn associations within large-scale datasets. However, since it is challenging to exert control over the neural generation process and the corresponding output, the usability of such models in practical scenarios are limited, as the generated content could be erroneous, incoherent, or even socially inappropriate (Liu et al., 2020; Wiseman et al., 2017). It is therefore ideal to include explicit provisions in neural text generation to better model characteristics such as informativeness and topical continuity. It has also been shown that informativeness and textual cohesion are important properties in clinical texts to make them more easily comprehensible (Smith et al., 2011; Liu and Rawl, 2012).

Image to text generation is a natural language generation task that has been popular in communities beyond NLP (e.g. computer vision, machine learning). A general approach is to construct the representation of the entire input image and decode the output text conditioned on the image representation (You et al., 2016). Such approaches work well for scenarios where only a short generated sentence is needed in the output (e.g. image captioning), as typically what is needed is to identify individual objects and fill in the most probable words to describe the overall situation. However, such approaches might not generalize to scenarios where complex semantics embodied in the input images need further inferencing or where the generated outputs need to articulate detailed or specific information, logical reasoning, or recommendations — all of these cases typically require at least multiple sentences (to form a report) (Jing et al., 2017). Medical reports are a classic example of such a scenario where each sentence in a report describes very precise clinical observations or inferences.

We present a neural approach for producing radiology reports from images in a sentence-by-sentence order to pinpoint more targeted and precise medical information from the input images and at the same time minimize hallucination from neural text generation. The modeling components ensures the generated report is informative, coherent, and concise via gated mechanisms to model topical continuity, orthogonality criteria in sentence state selection to reduce redundancy, and a neural architecture that is pretrained to predict domain entities during each context of sentence generation in order to encourage induction bias.

Proceedings of NAACL HLT 2021: IndustryTrack Papers, pages 246–254
June 6–11, 2021.

2 Related Work

2.1 Natural Language Generation

Quests for more efficient methods arising from machine translation using dense sentence representations resulted in the development of neural text-to-text generation models (Bahdanau et al., 2014; Cho et al., 2014; Srivastava et al., 2014; Wiseman et al., 2018). Subsequently, neural approaches for text-to-text generation for summarization tasks also started to gain traction (Cheng and Lapata, 2016; Nallapati et al., 2017; See et al., 2017; Paulus et al., 2017). A major interest in the medical NLP community focuses on information extraction (see Wang et al. (2018) for a review). There has been work in areas such as automatic ICD code assignment (Zhang et al., 2017; Scheurwegs et al., 2017; Mullenbach et al., 2018), risk prediction (Ma et al., 2018), and dialogue comprehension (Liu et al., 2019), and text generation (Buchanan et al., 1995; Moradi and Ghadiri, 2018; Pauws et al., 2019).

2.2 Image to Text

There has been much work in image to text generation, which typically constructs a representation of the input image using CNN and generates the output text using RNN (Fang et al., 2015; Krause et al., 2017; Vinyals et al., 2015). Such work has been improved further by incorporating the attention mechanism on input representations (Xu et al., 2015; You et al., 2016). Xu et al. (2015) used visual spatial attention for improving text generation while You et al. (2016) introduced semantic attention on concepts. All of the aforementioned work demonstrated effectiveness on single sentence generation such as captions. Image to text generation becomes more challenging when considering multi-sentence outputs. Some recent work generated multi-sentence outputs using hierarchical decoding (Krause et al., 2017; Liang et al., 2017). Jing et al. (2017) adapted this approach for radiology report generation by incorporating co-attention. Yuan et al. (2019) further improved the design by incorporating concept prediction and leveraging the predicted concept for guiding generation. In our work, the network is pre-trained to predict context entities so that each sentence generation is implicitly guided by domain entities. In addition, our system explicitly models informativeness and topical continuity to improve coherence while reducing redundancy to increase factual correctness and readability.

3 Method

In this section we delineate the following: (1) The proposed neural architecture and the corresponding network computations; (2) How we pre-train the network to predict the context entities from each sentence representation using a multi-label classifier; (3) How we further train the neural architecture to decode the corresponding sentences from each sentence representation to form the report.

3.1 Neural Architecture

Each input to our network is a set of images S_I with different views of the chest from the same patient and an indication text Q, which is a short sentence or phrase describing the purpose of the radiology investigation (e.g. intense coughing)[1]. Figure 1 depicts the architecture of our neural model, consisting of components for image encoding, indication text encoding, image feature selection for informativeness, sentential content creation for topical continuity and redundancy reduction and for decoding individual sentences in the report. Before the network computations commence, content creation RNN is initialized with a zero vector hidden state. We elaborate each component in the following subsections.

3.1.1 Image Encoding and Sentential Content Creation

Our network is designed to generate the radiology report in a sentence-by-sentence manner from the input set of images, guided by the indication text. The sentence-by-sentence design allows the report generation to focus on specific and important details in the medical image and reduces possible pitfalls of hallunciation in neural text generation. The Image encoder is a ResNet152 network with pretrained weighs (He et al., 2016). Using the encoder, each of the image matrix i in the input image set is converted to $I_i \epsilon R^n$, as depicted on the left hand side of Figure 1. The network updates the image representations during each context of sentence generation. The network employs gates for informative content selection and topical continuity weighted by a control gate.

Informative Content Selection: The content selection gate is represented by the trapezium on the top of Figure 1. Gate gc selects the

[1] An example of indication text: `male with cough and rib pain`.

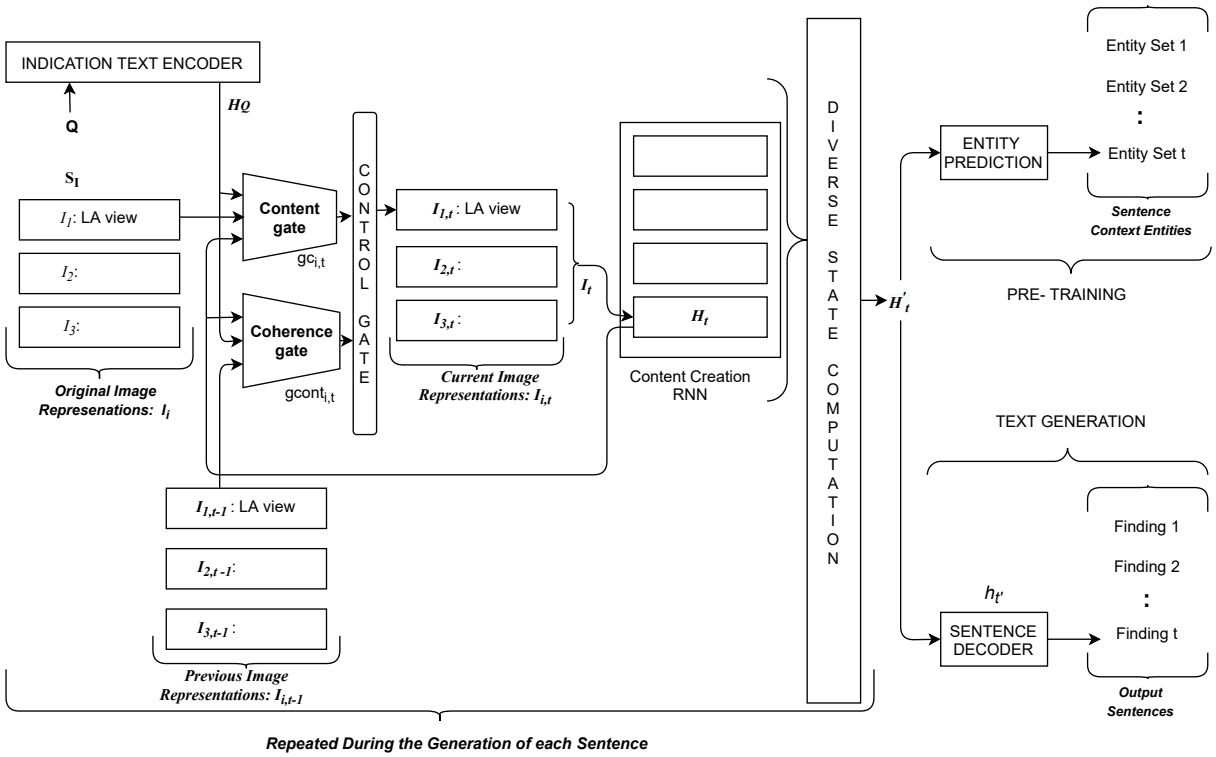

Figure 1: Proposed Neural Architecture

informative content from the original image representations during each time-step t of the content creation RNN. Gate gc filters the features of the input image representation I_i as follows:

$$gc_{i,t} = \text{sigmoid}(\mathbf{W_{gc}}[I_i; H_{t-1}; H_Q])$$
$$Ic_{i,t} = gc_{i,t} \odot I_i$$

where W_{gc} is the parameter matrix, H_{t-1} is the previous hidden state of content creation RNN and H_Q is the indication text encoded using a transformer network. The presence of H_{t-1} ensures that features are selected in the context of previously generated sentences.

Content Selection for Topical Continuity: The gate $gcont$ selects the content for topical continuity at time- step t from the image representations computed for the previous time- step $t-1$. In Figure 3.1, the continuity gate is represented by the trapezium at the bottom. The Gate $gcont$ selects the content for topical continuity as follows:

$$gcont_{i,t} = \text{sigmoid}(\mathbf{W_{gcont}}[I_{i,t-1}; H_{t-1}; H_Q]) \quad (1)$$

$$Icont_{i,t} = gcont_{i,t} \odot I_{i,t-1} \quad (2)$$

W_{gcont} is a parameter matrix and $I_{i,t-1}$ is the representation of the i^th image in the input set computed at time-step $t-1$.

Control Gate: The control gate is represented by the first vertical rectangle in Figure 1. Control gate weighs and creates the representation of the i^th image for time step t as follows.

$$\alpha_{i,t} = \text{sigmoid}(\mathbf{W_{cont}}[H_{t-1}]) \quad (3)$$
$$I_{i,t} = \alpha_{i,t} * Ic_{i,t} + (1 - \alpha_{i,t}) * Icont_{i,t}, \quad (4)$$

where W_{cont} is a parameter matrix.

3.1.2 Sentence Content Creation

The content creation RNN is represented by the vertical rectangle, encompassing smaller rectangles corresponding to different states as depicted in the middle of Figure 3.1. Content Creation RNN computes the content for the sentence to be decoded at time step t by taking final representations for the images in the input set into account. The input I_t at the current time step t of content creation RNN is computed as follows:

$$I_t = \frac{\sum_i I_{i,t}}{m} \quad (5)$$

where m is the number of images in the input set. The hidden state H_t for content creation RNN at time step t is computed as below.

$$H_t = \mathbf{GRU}(I_t, H_{t-1}), \quad (6)$$

3.1.3 Reducing Redundancy via Orthogonal Sentence States

Avoiding redundant content generation is a problem to be explicitly addressed by text generation systems (Nema et al., 2017). Hidden states of content creation RNN represents the content corresponding to each sentence in the final report. Enforcing diversity among these hidden representations can reduce the redundant content in the resultant report. We ensure that each hidden state of content creation RNN used to initialize decoder to be orthogonal to the mean of previous hidden states. In the purview of this orthogonality H_t of content creation RNN is updated as follows.

$$H_t^{'} = H_t - \frac{H_t^T H_{t-1}^M}{(H_{t-1}^M)^T H_{t-1}^M} \quad (7)$$

where H_{t-1}^M is the mean of previous hidden states.

3.2 Pre-Training via Entity Prediction

For the purpose of pre-training we predict context entities from the constructed content $H_t^{'}$ using a multi-label classifier:

$$\begin{aligned} H_t^{"} &= NN(H_t^{'}) \\ Scores_t &= \text{softmax}(H_t^{"}) \\ Ent_t^k &= arg\,max_k(Scores_t) \end{aligned} \quad (8)$$

NN is a two layered fully connected neural network where the individual layer computations are a linear transformation followed by a ReLU activation. Ent_t^k represents the set of top k ranking context entities, which are intended to contain the entities to be mentioned in the sentence to be generated at time-step t of content creation RNN. The pre- training is done using binary cross entropy loss.

3.3 Training the Sentence Decoder

We use a decoder with beam search decoding to generate sentences. The sentence decoder RNN is initialized with $H_t^{'}$, which represents the content to be materialized at time step t. At each time step $t^{'}$ of decoder RNN, a word in the sentence under construction is generated as follows:

$$P(w_{t'}|w_{<t'}) = \text{softmax}(\mathbf{W_o}(h_{t'}) + b_o), \quad (9)$$

where $h_{t'}$ is the hidden state of decoder RNN at time-step $t^{'}$. Negative log likelihood is used for training the network to generate sentences.

4 Results

4.1 Data Setup

A subset of 19,800 entries were selected from the MIMIC-CXR Database[2] for generating radiology reports from medical images of chest X-rays (Johnson et al., 2019), where each entry is represented by a triplet (S_I, Q, SEQ_F). S_I is a set of m input radiology images where there are one or more images corresponding to different views of a patient's chest, Q is a short text span specifying the purpose of the radiology investigation, and SEQ_F represents the sentences written by a radiologist in the context of S_I and Q. SEQ_F is a sequence of sentences $f_1, .., f_n$, each representing an individual finding.

We reformulate the dataset so that each entry record is (S_I, Q, SEQ_F, SEQ_E). SEQ_E represents a sequence of entity sets $ent_1, ..., ent_n$, where ent_i represents the set of entities mentioned in sentence f_i. We extracted entities from individual sentences[3] and identified a frequently occurring set of 1,060 entity clusters[4] suitable for learning to predict context entities and subsequent sentence generation. Sentences that do not consist a single mention of any of these entities were removed because they were evaluated to be subject to information not included in the corresponding images. Our dataset consists of 18,000, 900 and 900 training, test and development records respectively.

4.2 Experimental Setup

- **Img + RNN :** The entire radiology report is decoded as a single sequence from the mean of image representations (Fang et al., 2015).
- **Img + Attn :** The decoder RNN attends over the input image representation to generate a single sequence that constitutes the report (You et al., 2016).

[2] https://physionet.org/content/mimic-cxr/2.0.0/

[3] A pilot study was conducted to compare the effectiveness of entities extracted from https://spacy.io/ and https://ctakes.apache.org/, which showed no obvious difference between the two named entity recognition tools. The former was thus chosen due to ease of integration into our existing codebase.

[4] Entities that refer to the same medical phenomenon (e.g. *acute pneumonia* and *pneumonia*) were clustered to further streamline the modeling process.

Experimental Setting	Text Gen				Cont Order		
	R -1	R - 2	B -2	B - 4	P	R	F
Img + RNN	0.241	0.070	0.200	0.070	0.040	0.050	0.045
Img + Attn	0.270	0.079	0.200	0.075	0.060	0.070	0.064
Img + Pred + Co-Attn	0.291	0.093	0.250	0.091	0.081	0.110	0.093
Img + Ent + Attn	0.300	0.096	0.273	0.102	0.080	0.100	0.089
Img + IC	0.291	0.090	0.261	0.100	0.070	0.090	0.772
Img + IC + TC	0.310	0.097	0.291	0.109	0.090	0.126	0.100
Img + IC + TC + O	0.318	**0.109**	0.328	0.117	**0.120**	0.135	**0.127**
Img + IC+ TC + O + PT	**0.323**	0.106	**0.334**	**0.120**	0.117	**0.137**	0.126

Table 1: Report Generation Performance Comparison. Text Generation: R-1, R-2, B-2, B-4 denote ROUGE-1, ROUGE-2, BLEU-2, BLUE-4, respectively. Content Ordering: P: Precision, R:Recall, F: F-Measure.

Experimental Setting	HR	
	mean	std
Img + RNN	2.70	2.07
Img + Attn	3.51	1.70
Img + Pred + Co-Attn	6.16	1.46
Img + Ent + Attn	6.02	1.53
Img + IC	4.50	1.65
Img + IC + TC	6.10	1.45
Img + IC + TC + O	**7.02**	1.26
Img + IC+ TC + O + PT	6.71	1.33

Table 2: Human Ratings is denoted as HR; the mean and standard deviation (denoted as std) are computed for each setting, and the overall Pearson Coefficient is 0.67.

- **Img+ Pred + Co-Attn :** A multi-image variant of the co-attention based method (Jing et al., 2017), in which sentence context vectors by co-attending over input images and entities.
- **Img + Ent+Attn :** This setting is a variant of (Yuan et al., 2019), where the decoder attends over a predicted set of entities to generate sentences.
- **Our Method:** We experiment with different settings of our approach depicted in Figure 3.1 with different combinations of Informative Content selection **(IC)**, Topical Continuity **(TC)**, Orthogonal Sentence States **(O)** and Pre-Training **(PT)**.

Encoded image size is 900 after linear transformation of ResNet output, $H_t \epsilon R^{900}$ and $h_{t'} \epsilon R^{900}$. Other parameters are adjusted accordingly. For all settings, a beam size of 9 is set for the decoder. For all the settings and for each of the test record we generate five sentences as the average number of sentences in development set reports is approximately five. The set of parameters which gave maximum recall for entity prediction in the development set during pre- training is used initialize the network during training.

4.3 Text Generation and Content Ordering

We evaluated the quality of text generation using BLEU and ROUGE metrics as shown in Table 1. The setting $Img + IC$ did not perform well with respect to other counterparts. This suggests that just informative content selection gate and hidden state of content ordering RNN alone is insufficient for defining the context of a sentence. However $Img + IC + TC$ achieves an incremental accuracy by employing the efficient gated mechanism for sentence content creation. $Img + IC + TC + O$ performs consistently well on all metrics, especially using BLEU-4, implying the approach of eliminating redundant content in long text generation via enforcing topic diversity with orthogonal sentence states is effective. The setting with pre-training on entity prediction ($Img + IC + TC + O + PT$) achieved a slight incremental improvement in accuracy. We observe that for a large set of 1,060 domain entities, our training data is not dense enough for a significant improvement through pre-training. However the incremental improvement is encouraging.

Coherent reading results from accurate content ordering. For evaluating content ordering we relied on the method used by Kurisinkel and Chen (2019). They utilize the bigrams constituted by words in preceding and succeeding sentences irrespective of their positions within text in order to measure

ImgEnc + Ent + Attn
1) The lungs are clear without airspace consolidation.
2) Lungs are hyperinflated with no pleural effusion or pneumothorax is seen.
3) Lungs are hyperinflated with no pleural effusion or pneumothorax is seen.
4) The lungs are clear without focal consolidation.
5) No evidence of pleural effusion, pulmonary edema, pneumothorax, or focal airspace consolidation.
Img + IC+ TC
1) The lungs are clear without airspace consolidation.
2) No pleural effusion, pulmonary edema, pneumothorax, or focal consolidation.
3) Lungs are hyperinflated with no pleural effusion or pneumothorax is seen.
4) Lungs are hyperinflated with no pleural effusion or pneumothorax is seen.
5) Degenerative changes of the thoracic spine with calcification of the anterior longitudinal ligament are present.
Img + IC + TC + O + PT (Proposed)
1) The lungs are clear without focal consolidation.
2) Lungs are hyperinflated with no pleural effusion, pulmonary edema or pneumothorax is seen.
3) Interstitial prominence is chronic.
4) The cardiac and mediastinal silhouettes are stable.
5) Degenerative changes of the thoracic spine with calcification of the anterior longitudinal ligament are present.
Radiologist Report Written by Physicians
1) PA and lateral views of the chest demonstrate the lungs are well expanded, with no evidence of pleural effusion, pulmonary edema, pneumothorax, or focal airspace consolidation.
2) Mild interstitial prominence is chronic, and unchanged.
3) Previously demonstrated bilateral fat-containing Bochdalek hernias are better assessed on prior CT of the chest.
4) The heart is mildly enlarged, Otherwise, the cardiomediastinal silhouette is unremarkable.
5) Multilevel degenerative changes are noted throughout the thoracic spine, with calcification. of the anterior longitudinal ligament

Table 3: Text generated from different experimental setups. Capital letters and dots are manually added for ease of reading. Red text: redundancy; blue text: named entities; green text: content that cannot be inferred from the given medical images in the dataset.

the accuracy of ordering. Accuracy depends on the overlap of such bigrams in generated and ref-

erence texts, measured using Precision, Recall and F- Measure. The results of content ordering are shown on the right side of Table 1. It is evident that adding explicit means for topical continuity (TC) and Redundancy reduction (O) increased the quality of content ordering at each phase.

4.4 Human Evaluation

We resort to human evaluation for rating the factual accuracy of radiology reports with respect to the reference report in hand. Four human evaluators were asked to rate the reports generated by all settings in Table 2 for a set of 100 test records that were randomly chosen. Reports were presented to the evaluators in a random order to minimize potential bias. The rating of a sentence is the sum of individual ratings of all the sentences in a report. Sentences describing an abnormal condition is weighed more than a sentence explaining a normal condition as they are clinically more relevant. A non-redundant sentence explaining an accurate normal condition is given a rating of 1.5 while that explaining an abnormal condition is given a rating of 3. A factually incorrect or redundant sentence receives a score of 0. The mean and standard deviation for each experimental setting are shown in the rightmost column of Table 1. The Pearson Coefficient is 0.67, suggesting that the agreement among the human evaluators are reasonably consistent (Benesty et al., 2009). The settings with content selection and continuity gates and diverse state computation achieved a clear advantage over the other settings implying it is effective to generate specific content for each sentence while explicitly eliminating redundancy in our proposed approach.

4.5 Qualitative Comparisons

Examples of radiology reports generated by different settings for the same set of images are shown in Table 3 to give readers more qualitative context of the generation results. Settings which used the gated mechanism for sentence content creation and orthogonal state computation better emulate human written reports in terms of informativeness and content ordering. There is an adequate number of domain entities in the generated report. which are found to be clinically relevant when compared with the corresponding human written report. There are portions of text in the human written report which are subjective to the situation and are irrelevant in the objective scheme of text generation.

5 Conclusion

We presented a technical approach on radiology report generation which ensures global text properties such as informativeness, topical continuity for coherence while reducing redundant content. Both objective metrics and human evaluations showed significant performance over competitive baselines.

6 Acknowledgements

The authors would like to thank the insightful discussions with I. Ho Mien, Z. Liu, P. Krishnaswamy, M. Nguyen, R. Puduppully, B. Unnikrishnan, and Y. Zhang. This research was supported by grant funding from A*STAR, Singapore (CR-2020-001, SSF A1818g0044, IAF H19/01/a0/023).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

Bruce G Buchanan, Johanna D Moore, Diana E Forsythe, Giuseppe Carenini, Stellan Ohlsson, and Gordon Banks. 1995. An intelligent interactive system for delivering individualized information to patients. *Artificial intelligence in medicine*, 7(2):117–154.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494.

Kyunghyun Cho, B van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8), 2014*.

Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back.

In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Baoyu Jing, Pengtao Xie, and Eric Xing. 2017. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*.

Alistair EW 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 de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6.

Philip Kinghorn, Li Zhang, and Ling Shao. 2018. A region-based image caption generator with refined descriptions. *Neurocomputing*, 272:416–424.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text generation from knowledge graphs with graph transformers. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2284–2293.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–325.

Litton J Kurisinkel and Nancy Chen. 2019. Set to ordered text: Generating discharge instructions from medical billing codes. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6166–6176.

Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P Xing. 2017. Recurrent topic-transition gan for visual paragraph generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3362–3371.

Chiung-Ju Liu and Susan M Rawl. 2012. Effects of text cohesion on comprehension and retention of colorectal cancer screening information: A preliminary study. *Journal of health communication*, 17(sup3):222–240.

Haochen Liu, Zhiwei Wang, Tyler Derr, and Jiliang Tang. 2020. Chat as expected: Learning to manipulate black-box neural dialogue models. *arXiv preprint arXiv:2005.13170*.

Zhengyuan Liu, Hazel Lim, Nur Farah Ain Suhaimi, Shao Chuen Tong, Sharon Ong, Angela Ng, Sheldon Lee, Michael R Macdonald, Savitha Ramasamy, Pavitra Krishnaswamy, et al. 2019. Fast prototyping a dialogue comprehension system for nurse-patient conversations on symptom monitoring. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 24–31.

Fenglong Ma, Jing Gao, Qiuling Suo, Quanzeng You, Jing Zhou, and Aidong Zhang. 2018. Risk prediction on electronic health records with prior medical knowledge. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1910–1919. ACM.

Milad Moradi and Nasser Ghadiri. 2018. Different approaches for identifying important concepts in probabilistic biomedical text summarization. *Artificial intelligence in medicine*, 84:101–116.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Preksha Nema, Mitesh Khapra, Anirban Laha, and Balaraman Ravindran. 2017. Diversity driven attention model for query-based abstractive summarization. *arXiv preprint arXiv:1704.08300*.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Steffen Pauws, Albert Gatt, Emiel Krahmer, and Ehud Reiter. 2019. Making effective use of healthcare data using data-to-text technology. In *Data Science for Healthcare*, pages 119–145. Springer.

Elyne Scheurwegs, Boris Cule, Kim Luyckx, Léon Luyten, and Walter Daelemans. 2017. Selecting relevant features from the electronic health record for clinical code prediction. *Journal of biomedical informatics*, 74:92–103.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Catherine Arnott Smith, Scott Hetzel, Prudence Dalrymple, and Alla Keselman. 2011. Beyond readability: investigating coherence of clinical text for consumers. *Journal of Medical Internet Research*, 13(4):e104.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. 2018. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2018. Learning neural templates for text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3174–3187.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659.

Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. 2019. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 721–729. Springer.

Danchen Zhang, Daqing He, Sanqiang Zhao, and Lei Li. 2017. Enhancing automatic icd-9-cm code assignment for medical texts with pubmed. In *BioNLP 2017*, pages 263–271.