7-11-2022

Improving the Factual Accuracy of Abstractive Clinical Text Summarization using Multi-Objective Optimization

Amanuel Alambo
Wright State University - Main Campus

Tanvi Banerjee
Wright State University - Main Campus, tanvi.banerjee@wright.edu

Krishnaprasad Thirunarayan
Wright State University - Main Campus, t.k.prasad@wright.edu

Mia Cajita

Improving the Factual Accuracy of Abstractive Clinical Text Summarization using Multi-Objective Optimization

Amanuel Alambo[1], Tanvi Banerjee[1], Krishnaprasad Thirunarayan[1], and Mia Cajita[2]

[1]Wright State University
[2]University of Illinois, Chicago

***Abstract*—While there has been recent progress in abstractive summarization as applied to different domains including news articles, scientific articles, and blog posts, the application of these techniques to clinical text summarization has been limited. This is primarily due to the lack of large-scale training data and the messy/unstructured nature of clinical notes as opposed to other domains where massive training data come in structured or semi-structured form. Further, one of the least explored and critical components of clinical text summarization is factual accuracy of clinical summaries. This is specifically crucial in the healthcare domain, cardiology in particular, where an accurate summary generation that preserves the facts in the source notes is critical to the well-being of a patient. In this study, we propose a framework for improving the factual accuracy of abstractive summarization of clinical text using knowledge-guided multi-objective optimization. We propose to jointly optimize three cost functions in our proposed architecture during training: *generative loss, entity loss and knowledge loss* and evaluate the proposed architecture on 1) clinical notes of patients with heart failure (HF), which we collect for this study; and 2) two benchmark datasets, Indiana University Chest X-ray collection (IU X-Ray), and MIMIC-CXR, that are publicly available. We experiment with three transformer encoder-decoder architectures and demonstrate that optimizing different loss functions leads to improved performance in terms of entity-level factual accuracy.**

***Index Terms*—Clinical Text Summarization, Multi-Objective Optimization, Transformers, Heart Failure, Named Entity Recognition, Knowledge Bases, Factual Accuracy**

I. Introduction

Recent advances in transformer-based models [1] have led to progress in abstractive summarization of news articles, scientific articles, and social media data. However, these models have not been well investigated in the healthcare domain where automated clinical summary generation [2] for a set of findings in clinical notes is helpful to clinicians in saving their time and improving clinical decision making. One of the clinical practices by medical professionals entails the task of recording *findings* of diagnosis, treatment or procedures followed by summarizing the findings into a form called *impression*. Inspired by recent efforts in modeling findings-to-impression as summarization [3]–[5], we propose to automate this process of writing an impression for findings to assist clinicians with their practice, making the clinical workflow more efficient. Specifically, we attempt to accomplish this using an abstractive approach to summarizing findings into an impression. Further, clinicians use their commonsense understanding and their knowledge of the domain while producing an impression in addition to what is explicitly stated in the findings. As such, an impression has to be factually correct with respect to the findings. This is particularly critical in the healthcare setting where a misinterpreted impression could prove fatal and should be avoided at all costs to deliver quality health to patients. This issue is further exacerbated in the sub-domain of *Heart Failure* (HF) where reliable diagnosis is challenging and the cost of inaccuracy can be enormous. To investigate these issues, we utilize clinical notes of 1200 patients with HF from the University of Illinois Hospital & Health Sciences System (UI Health) for our study. Figure-1 illustrates what a typical clinical note (a record) for a patient with HF in our cohort looks like.

In this paper, we model and automate this clinician's impression (summary) writing process using a two-stage approach: 1) clinical knowledge retrieval from domain-specific knowledge sources using named entities; and 2) joint training of end-to-end transformer encoder-decoder models using multi-objective optimization. We evaluate our proposed framework on two benchmark datasets and a dataset we prepare for this task (heart-failure - HF) and demonstrate that we achieve significantly better results over baseline model training settings of *findings-to-impression* on factual accuracy metrics. We use three SOTA pre-trained transformer-encoder-decoder networks and fine-tune them on our datasets using different training objectives and report the results of the fine-tuning for each dataset. Our experimental data for HF consists of 6182 patient records where each record

Research reported in this publication was supported by NCCIH of the National Institutes of Health under award number 5R01AT010413S1.

Corresponding author: Amanuel Alambo (alambo.2@wright.edu)

Procedure_name:
[PERSONALNAME] Abd and Pelv w/o [PERSONALNAME] cont

Indication:
64-year-old female with history of incarcerated hernia, concern for small bowel obstruction

---

Technique:
Multidetector multiplanar noncontrast [PERSONALNAME] images through the abdomen and pelvis were obtained.

Comparison:
[PERSONALNAME] examination of the abdomen and pelvis

Findings:
Lack of intravenous contrast limits exam interpretation. LUNG BASES: There is a moderate right pleural effusion. There is dependent atelectasis in the lung bases. The heart is slightly increased in size compared to prior examination. There is atherosclerotic calcification of the coronary arteries. LIVER: The liver demonstrates cirrhotic morphology with nodular surface contours. GALLBLADDER AND BILIARY SYSTEM: There are no calcified gallstones. SPLEEN: The spleen is borderline enlarged measuring up to 13 cm in length. PANCREAS: Evaluation of the pancreas is suboptimal in the absence of [PERSONALNAME] contrast. ADRENAL GLANDS: There is a low attenuating lesion in the left adrenal gland measuring approximately 2.2 cm (series 2 image 22) that appears slightly enlarged since February 27, 2018 when it measured approximately 1.9 cm. This is favored to represent an adenoma. KIDNEYS: In the inferior pole cortex of the left kidney there is a 1.2 cm simple cyst, more conspicuous than the prior study. STOMACH: The stomach is mildly distended with air and debris. BOWEL: Postsurgical changes with bowel sutures are again seen in the right lower quadrant. There is mild small bowel dilatation adjacent to the suture line, probably within normal limits postsurgical. There is a focally dilated loop of small bowel in the left mid abdomen measuring up to 4.2 cm (series 2 image 30) with passage of oral contrast distally, suspicious for partial small bowel obstruction. There is passage of oral contrast to the level of the terminal ileum. There is amorphous soft tissue in the mid abdomen (series 2 images 53, 54) which likely represents unopacified small bowel loops rather than mass. There is scattered stool in the colon. PERITONEUM AND RETROPERITONEUM: There is mild to moderate volume ascites. There is no intraperitoneal free air. The abdominal aorta is normal in course and caliber with atherosclerotic calcifications throughout its abdominal course extending into the common iliac arteries.There is no mesenteric or retroperitoneal lymphadenopathy. PELVIS: The urinary bladder is well distended and unremarkable. There is no pelvic lymphadenopathy. Multiple phleboliths are again seen. BONES: There are mild degenerative changes of the spine with a diffuse disc bulges at L4-L5 and L5-S1. Sclerosis of L4-L5 appears unchanged since prior examination. SOFT TISSUES: There is anasarca in the subcutaneous soft tissues. A midline laparotomy scar is again seen.

Impression:
1. Findings suspicious for a proximal, partial small bowel obstruction.
2. Moderate right pleural effusion.
3. Cirrhotic liver morphology.
4. Moderate volume ascites.
5. Postsurgical changes of small bowel resection in the right lower quadrant.
6. Slight increase in size of left adrenal nodule favored to represent an adenoma. These images were reviewed and interpreted with attending radiologist Dr. [PERSONALNAME] before dictation of this final report by resident Dr. [PERSONALNAME].

Fig. 1: Example de-identified clinical record for the heart failure data collected through the Center for Clinical and Translational Science, University of Illinois, Chicago.

comprises *Procedure Type, Techniques, Indication, Findings* and *Impression* tuples of clinical notes.

The main contributions of this study are: 1) the introduction of an approach for clinical named entity-aware knowledge retrieval from medical knowledge sources; 2) a novel training technique for abstractive summarization of clinical text using multi-objective optimization; and 3) new experiments using the state-of-the-art transformer encoder-decoder networks as backbone models to demonstrate that optimizing knowledge-driven cost functions in addition to the generative cost function during training boosts model performance on factual accuracy metrics.

II. Related Work

The birth of Transformer encoder-decoder models [1], [6]–[9] has led to significant advances in abstractive summarization in the domains of news articles [10]–[12], and scientific articles [13]–[15]. Nevertheless, their application to the summarization of clinical notes has not been adequately explored. [5] proposed a model based on Pointer-Generator-Networks [12] for abstractive summarization of radiology reports by linking entities in a clinical note to domain-specific ontology from UMLS [16] and RadLex [17]. They use pairings of findings and impression for the abstractive summarization task where findings form the input sequences and impressions form the target summaries for training. [18] propose a two-stage model consisting of a content selector and abstractive summarizer for clinical abstractive summarization. The content selector identifies ontological terms from the findings using a medical ontology (RadLex) and the summarizer is trained to generate summaries (impressions). They use Bi-LSTMs to encode findings and use LSTMs to encode the ontological terms followed by an LSTM-based decoder to generate a summary. [19] built a model for extractive summarization of clinical notes of patients with diabetes and hypertension to generate disease-specific summaries. They framed the summary generation problem as a sentence classification problem and experimented on a dataset consisting of 3,453 clinical notes collected for 762 patients. [20] proposed a model comprised of syntax-based negation detection and semantic clinical concept recognition for extractive summarization of clinical text. They conducted their experiments on the MIMIC-III [21] dataset. While the aforementioned approaches employ different techniques for clinical text summarization, we show experimentally that our proposed knowledge-aware Multi-Objective Optimization (MOO) improves the factual accuracy of the generated summaries when compared to strong state-of-the-art transformer-based abstractive summarization models.

III. Data Preparation

Out of the total of 15183 de-identified procedure notes spanning a period of over 4 years (5/2016 - 8/2020) collected from patients with HF admitted to UI Health, we filter the ones with no Findings or Impressions since our task is to generate an impression for a finding. Thus, the findings play the role of input text to be summarized and the impression serves as the ground truth summary. After pre-processing the data, we have 6182 notes consisting of findings-to-impression pairings along with other metadata. In addition to our Heart Failure data, we evaluate the proposed approach on two benchmark datasets on radiology reports from the Indiana Network for Patient Care [22] and 50000 randomly selected chest x-ray reports from the MIMIC-III-CXR dataset [23].

IV. Proposed Approach

A. Clinical Text Named Entity and Knowledge Extraction

We use an off-the-shelf Stanza package from Stanford for clinical named entity recognition (NER) [24]. Specifically, the Stanza model we use is the one trained on the i2b2 clinical text dataset. The knowledge bases to query for facts using the named entities are composed of UMLS, SNOMED-CT, and

ICD-10. After named entities are extracted using Stanza from a finding, our next task is to query for facts pertaining to the named entities as they appear in domain-specific knowledge bases. For each named entity identified from a finding, we perform full-text lexical query of the KBs and return the top-k facts where we set the value of K to 5 [25].

B. Model Training using Multi-Objective Optimization

We experiment with three state-of-the-art transformer-based models pretrained using different self-supervised objectives. We propose to train these models using a loss function that optimizes summary generation, named entity chain generation, and fact generation where our task is not only to auto-regressively generate the target summary, but also to generate the named entities in the impression and to generate the facts associated with the named entities in the impression. Figure-2 shows the proposed end-to-end architecture where three networks, whose parameters are shared are jointly trained using the loss functions stated in Equation-1.

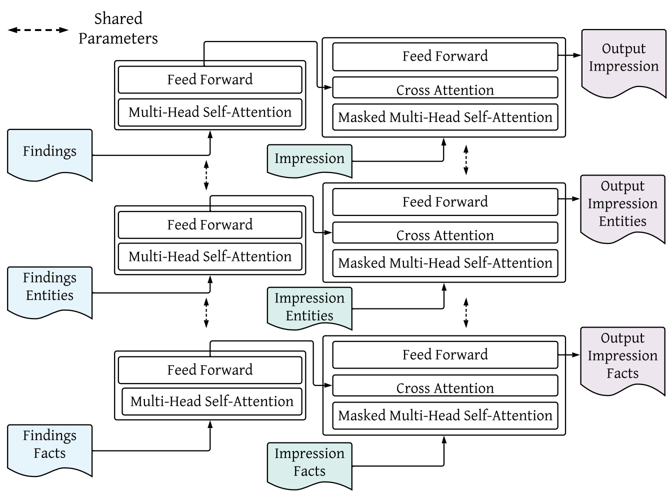

Fig. 2: The proposed training architecture.

We optimize the total aggregate loss function during the training phase for the proposed model in use. We use Bayesian optimization [26] to search for the best combination of *generative* and *regularization* hyperparameters. The generative hyperparameter is denoted in the formulation using λ_{gen} while the knowledge and entity-based regularization hyperparameters are denoted using λ_k, $\lambda_{\mathcal{E}}$. Each of the hyperparameters takes on values in the range of [0.1, 0.9] with increments of 0.3 and we evaluate the validation loss in each epoch during training to save the model checkpoint with the least validation loss. We experiment with three optimization configurations: i) with generative loss alone; ii) with generative loss and entity chain loss; iii) with generative loss, knowledge loss, and entity chain loss.

TABLE I: Statistics of the experimental datasets.

Dataset	Train	Validation	Test	Avg # tokens per Findings	Avg # tokens per Impression
Heart Failure (HF)	4000	1091	1091	142	48
IU X-Ray	2200	593	593	33	12
MIMIC-CXR	40000	5000	5000	52	18

$$\mathcal{L}_{total} = \lambda_{gen} \cdot \mathcal{L}_{gen} + \lambda_k \cdot \mathcal{L}_k + \lambda_{\mathcal{E}} \cdot \mathcal{L}_{\mathcal{E}} \quad (1)$$

Each of the loss functions is based on cross-entropy criterion.

$$\mathcal{L}_{\theta} = -\frac{1}{n} \sum_{k=1}^{n} \mathcal{P}(t_k | t_{<k}, \chi; \theta) \quad (2)$$

Where χ - the input sequence (i.e., finding, or named entity chain in a finding, or a sequence of facts retrieved from the knowledge bases associated with named entities in a finding). The proposed models are trained with the objective of minimizing the aggregate loss function defined in Equation-1. All models are built and trained using PyTorch on Google Cloud NVIDIA Tesla T4 GPU.

V. Experimental Results and Discussion

Table-I shows the statistics of the datasets and Table-II shows the results of evaluation against the impressions (ground truth summary). Our experimental results show that jointly optimizing the task of traditional language modeling with task-specific objectives such as preserving entity-aware factual accuracy improves performance of a model. Specifically, we demonstrate this by leveraging three pre-trained abstractive summarization models and fine-tuning on our datasets using multi-objective optimization. As can been from Table-II, Precision-target, and Recall-target increase with our training objective as compared to the language modeling training objective used with the baseline models. As extensively discussed in the literature [4], [27], we also argue that lexical measures (i.e., ROUGE) do not fully quantify the factual accuracy of a generated summary while a metric that measures entity-level overlap between a ground truth summary (impression) and a model-generated summary better reflects the extent to which semantics are preserved in abstractive summarization since named entities constitute significant semantics in a clinical text. A key limitation of our proposed approach is it is computationally more expensive and takes longer to train than with customary single task objective training. Another limitation we observed is that the proposed model training approach can be sensitive to hyperparameter initialization.

TABLE II: Experimental results. Dual MOO refers to dual multi-objective optimization where only the generative loss and entity chain loss are jointly optimized during training. Triple MOO refers to modeling where the three loss functions are jointly optimized. Due to space constraints, we report average scores across the three datasets.

Model	R-1	R-2	R-L	Entity-level Factual Accuracy		
				Precision-target	Recall-target	F1 score-target
T5 Vanilla (Baseline)	**35.113**	**19.503**	**34.921**	25.150	42.577	31.621
T5 w/ named entities (dual MOO) - Ours	32.628	18.361	33.827	**29.672**	46.581	36.252
T5 w/ named entities /w facts (triple MOO) - Ours	28.761	17.382	30.599	29.327	**48.148**	**36.451**
BART Vanilla (Baseline)	**22.951**	**16.283**	**22.657**	18.321	29.679	22.656
BART w/ named entities (dual MOO) - Ours	19.827	13.693	19.792	20.629	33.839	25.632
BART w/ named entities /w facts (triple MOO) - Ours	15.721	12.173	16.582	**23.182**	**34.159**	**27.620**
Pegasus Vanilla (Baseline)	**28.193**	**11.387**	**28.079**	21.739	28.593	24.699
Pegasus w/ named entities (dual MOO) - Ours	27.370	9.728	25.372	22.058	**29.781**	**25.344**
Pegasus w/ named entities /w facts (triple MOO) - Ours	24.263	7.836	22.174	**25.661**	25.349	25.504

VI. Conclusion and Future Work

In this study, we proposed a framework based on a transformer encoder-decoder network and transfer learning for clinical text summarization using knowledge-aware multi-objective optimization. We experimentally demonstrated that jointly optimizing generative loss, knowledge loss, and entity-based loss functions significantly improves the quality of generated summaries in terms of entity-level factual accuracy which is critical but less explored in the healthcare domain. In the future, we plan to extend the proposed multi-task learning framework for a different healthcare domain. Further, while the current study utilizes standard cross-entropy for each loss function, we plan to experiment with different loss functions including other entropy-based functions (e.g., KL-divergence) for the regularization components. In addition, while the knowledge retriever in the proposed approach is an independent unit from the summarizer, we plan to extend the proposed end-to-end training framework to include the knowledge retriever as one component of the framework.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[2] R. Pivovarov and N. Elhadad, "Automated methods for the summarization of electronic health records," *Journal of the American Medical Informatics Association*, vol. 22, no. 5, pp. 938–947, 2015.

[3] Y. Zhang, D. Y. Ding, T. Qian, C. D. Manning, and C. P. Langlotz, "Learning to summarize radiology findings," *arXiv preprint arXiv:1809.04698*, 2018.

[4] Y. Zhang, D. Merck, E. B. Tsai, C. D. Manning, and C. P. Langlotz, "Optimizing the factual correctness of a summary: A study of summarizing radiology reports," *arXiv preprint arXiv:1911.02541*, 2019.

[5] S. MacAvaney, S. Sotudeh, A. Cohan, N. Goharian, I. Talati, and R. W. Filice, "Ontology-aware clinical abstractive summarization," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 1013–1016.

[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *arXiv preprint arXiv:1910.10683*, 2019.

[8] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 328–11 339.

[7] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[9] W. Qi, Y. Yan, Y. Gong, D. Liu, N. Duan, J. Chen, R. Zhang, and M. Zhou, "Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training," *arXiv preprint arXiv:2001.04063*, 2020.

[10] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," *Advances in neural information processing systems*, vol. 28, pp. 1693–1701, 2015.

[11] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," *arXiv preprint arXiv:1509.00685*, 2015.

[12] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," *arXiv preprint arXiv:1704.04368*, 2017.

[13] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian, "A discourse-aware attention model for abstractive summarization of long documents," *arXiv preprint arXiv:1804.05685*, 2018.

[14] I. Cachola, K. Lo, A. Cohan, and D. S. Weld, "Tldr: Extreme summarization of scientific documents," *arXiv preprint arXiv:2004.15011*, 2020.

[15] M. Yasunaga, J. Kasai, R. Zhang, A. R. Fabbri, I. Li, D. Friedman, and D. R. Radev, "Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 7386–7393.

[16] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[17] C. P. Langlotz, "Radlex: a new method for indexing online educational materials," pp. 1595–1597, 2006.

[18] S. Sotudeh, N. Goharian, and R. W. Filice, "Attend to medical ontologies: Content selection for clinical abstractive summarization," *arXiv preprint arXiv:2005.00163*, 2020.

[19] J. Liang, C.-H. Tsou, and A. Poddar, "A novel system for extractive clinical note summarization using ehr data," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 2019, pp. 46–54.

[20] W.-H. Weng, Y.-A. Chung, and S. Tong, "Clinical text summarization with syntax-based negation and semantic concept identification," *arXiv preprint arXiv:2003.00353*, 2020.

[21] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[22] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 304–310, 2016.

[23] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.

[24] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A python natural language processing toolkit for many human languages," *arXiv preprint arXiv:2003.07082*, 2020.

[25] C. An, M. Zhong, Z. Geng, J. Yang, and X. Qiu, "Retrievalsum: A retrieval enhanced framework for abstractive summarization," *arXiv preprint arXiv:2109.07943*, 2021.

[26] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," *Advances in neural information processing systems*, vol. 25, 2012.

[27] B. Goodrich, V. Rao, P. J. Liu, and M. Saleh, "Assessing the factual accuracy of generated text," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 166–175.