INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Investigating the Utility of Synthetic Data for Doctor-Patient Conversation Summarization

Siyuan Chen[1*], *Colin A. Grambow*[2*], *Mojtaba Kadkhodaie Elyaderani*[2], *Alireza Sadeghi*[2], *Federico Fancellu*[2], *Thomas Schaaf*[2]

[1]University of Illinois at Urbana-Champaign
[2]3M Health Information Systems

siyuanc2@illinois.edu, colingrambow@gmail.com,
{mkadkhodaieelyaderani, asadeghi, ffancellu, tschaaf}@mmm.com

Abstract

Large-scale pre-training has been a successful strategy for training transformer models. However, maintaining a large clinical dataset for pre-training is not always possible, and access to data in this domain can be time-limited and costly. We explore using synthetic data in pre-training sequence-to-sequence (seq-to-seq) transformer models to generate clinical notes from Doctor-Patient-Conversations (DoPaCos). Using a generative language model fine-tuned on authentic conversations, a synthetic DoPaCo dataset was created and used with a corpus of clinical notes to pre-train a Longformer-Encoder-Decoder (LED) model. Results show that synthetic data leads to comparable performance in the downstream summarization task compared to pre-training with authentic data. Pre-training on synthetic conversations first, followed by clinical notes, yields higher performance across most of our evaluation metrics.

Index Terms: Natural Language Generation, Data Augmentation, Pre-training, Doctor Patient Conversation Summarization.

1. Introduction

Clinical notes written by U.S. physicians have nearly doubled in length since 2009 and are now nearly four times longer on average than those in other countries across the same electronic health records (EHR) system [1]. A new medical scribe industry has emerged to alleviate the additional workload associated with EHR use [2] and has been shown to improve physician satisfaction and quality of clinical notes [3]. However, this has merely shifted the burden from physicians to scribes, and continued reliance on human experts has resulted in substantial costs.

Many recent works seek to alleviate this burden on physicians and medical scribes by exploring the use of seq-to-seq Transformer models for automatically generating clinical notes from doctor-patient conversations (DoPaCos). A challenge commonly encountered in this task is that DoPaCos are typically lengthy and thus can easily violate the input limits of most pre-trained Transformer models. To circumvent this limitation, extract-then-abstract approaches have been introduced in [4, 5], where a classifier first filters out the irrelevant utterances for the summarization task, and then a pre-trained model, such as T5 [6] or BART [7], summarizes extracted text. An alternative approach is to advocate for a two-stage summarization scheme as in [8]. Others have employed end-to-end abstractive summarization methods and tackled the input length limits using specialized models designed for longer documents. A variety of seq-to-seq models have been investigated in [9], where authors pre-trained Transformer-based summarizers on medical conversations and evaluated their performance with automatic clinical concept overlap metrics. A similar work [10] has fine-tuned a summarizer model consisting of a Longformer encoder and a BERT decoder after pre-training it on an EHR dataset.

Some recent works on this topic have taken advantage of in-domain pre-training of general-purpose pre-trained models before fine-tuning them on the target task. However, licensed access to datasets in the medical domain, especially those that include sensitive patient information, is expensive and often time-limited. Following the trails of recent research in data augmentation using pre-trained language models [11, 12], [13] has attempted to augment the training data for the DoPaCo summarization task by using a GPT-3 ensemble model to generate summaries for unannotated DoPaCos. Although it is promising that certain models trained on synthetic summaries were outperforming those trained on human-generated ones, and that a clinical NER system was incorporated during the ensemble generation process to enhance medical concept coverage for the synthetic summaries, these models do not address the generation of DoPaCos, which are typically scarce.

To address this shortcoming, this paper investigates the use of synthetic DoPaCos in pre-training Transformer models for generating the History of Present Illness (HPI) section of clinical notes based on medical conversation transcripts. We first fine-tune a DialoGPT model [14] on a set of authentic, unlabeled DoPaCo transcripts and use it to generate a large number of synthetic DoPaCos. These synthetic conversations can then be used for in-domain pre-training of a Longformer-Encoder-Decoder (LED) model [15], before it is finally fine-tuned on a small set of annotated conversations for the downstream summarization task. We assess the performance of our summarization models using a variety of metrics, including N-gram overlap metrics, as well as semantic similarity and named entity-based metrics (Section 2.4). We show that synthetic conversations offer benefits comparable to authentic data when used for in-domain pre-training of LED models (Section 3.1). Our results also suggest that separated pre-training of encoder and decoder weights using either conversations or clinical notes may provide improvements compared to naively pre-training the model on all data in one pass (Section 3.2).

2. Methods

2.1. DoPaCo and clinical note datasets

We start with a dataset of roughly 80k manually transcribed and de-identified authentic DoPaCos. The conversations in this dataset lack reference summaries, which prevents us from using them for fine-tuning summarization models. Moreover, our access to this dataset is time-limited. We use this dataset to fine-

[*]Work done while at 3M Health Information Systems.

10.21437/Interspeech.2023-913

tune a DialoGPT dialogue response generation model[1], which is then used to generate about 160k synthetic DoPaCos (see Section 2.2). We intentionally generated more synthetic conversations because they are assumed to be of lower quality than real conversations, and it is hoped that increasing the quantity of data generated at little cost can compensate the decrease in quality. For fine-tuning of our DoPaCo summarization models, we use a separate set of 1342 annotated DoPaCos with corresponding human-written HPI summaries. The annotated set is split into 939/201/202 conversations for training, validation, and testing, respectively. The HPI notes have a mean length of 126 words and a standard deviation of 83 words.

In addition, we have access to a large corpus of proprietary, de-identified clinical notes spanning various specialties. About 475k of these notes were sampled from the corpus and used for pre-training of LED models. The clinical notes dataset contains complete SOAP[2] notes with a mean length of 374 words, standard deviation of 302 words, and minimum length of 50 words.

Best efforts have been made to de-identify the data, but it is possible that protected health information (PHI) still persists. Adding to the concern of possible PHI, the datasets used are proprietary and, due to contractual obligations cannot be shared.

Table 1: *An example of a generated conversation snippet.*

Prompt (excerpt from real conversation transcripts).
[DR]: Oh, okay. Yeah, so, you know, I mean, that's, you are doing everything you should do. Now, since, you know, you're, depression is a medicine resistant, I know you're on Prozac
Generated conversation.
[PT]: Um-hum. Yeah. I'm still on it. I just don't take it as often as I should. I know I need to take it more often. Um, I don't know if there's anything else that I can take that would help with the anxiety. I've been taking the Klonopin for a long time. [DR]: Yeah, I mean, you can take Ativan or lorazepam. Have you taken that before? Does that sound familiar? It's a little bit like Valium, but it's not really a Valium. It's, uh, it's a benzodiazepine. So it might help you to get through the day a little better. [PT]: Okay. I'll try it. [LAUGHING]. I'm willing to try anything. [INAUDIBLE]. I don't want to be drugged. I'm not that kind of person. I've never been drugged, but I've heard horror stories about it. So I just want to make sure that I'm, I'm alert and -. Okay.

2.2. Synthetic DoPaCo generation

We use a medium-sized DialoGPT model [14] with 345 million parameters to synthesize DoPaCos. The model is fine-tuned on 95 percent of our 80k unannotated DoPaCos for in-domain adaptation, with the remaining 5 percent of conversations left out for model validation. We set the learning rate to 2.5×10^{-5}, total training steps to 10^6 steps, training batch size to 8, and use a linear learning rate scheduler with 1000 warm-up steps.

To generate synthetic DoPaCos, we explore multiple generation heuristics, including beam search, top-k and top-p sampling combinations, and typical sampling [16]. For the final generation of synthetic data, we choose a combination of top-k (with $k = 100$) and top-p (with $p = 0.95$). This is because of their computational efficiency as well as the proximity of the word-distribution of generated DoPaCos to that of the authentic DoPaCos as measured by the Jensen-Shannon divergence.

The multi-turn conversation generation process is as follows. We first obtain a set of generic prompts from the authentic DoPaCos and pass a prompt with its speaker role label into the model. The model output is then appended to the end of the generated conversation string. To preserve the context of the conversation for the next generation round, we keep the last-generated 32 words as the new prompt, prepended with the corresponding speaker role label. We repeat the procedure until the generated conversation reaches the predefined maximum length of 800 tokens. We use a length penalty of 0.7 and a no-repeat N-gram size of 3. An example of a generated conversation snippet and its associated prompt is shown in Table 1.

2.3. Longformer summarizer pre-training and fine-tuning

We base all experiments on pre-training and fine-tuning of LED [15] model and initialize all models by the pre-trained weights publicaly available in the Transformers Library [17][3].

We adopt a DialogLED-style window-based denoising pre-training task [18] for both conversations and clinical notes, which for the LED-large model was found to show improvement in DoPaCo summarization tasks when pre-trained with authentic conversations [9]. We do not use sentence or turn permutation as they are found to be detrimental for the performance of the pre-trained models on the downstream task. When applied to clinical notes, the pre-training method can be considered as a basic windowed text denoising task, as clinical notes lack speaker labels that are relied on by the DialogLED-style turn splitting/merging pre-training tasks.

Using these models, we first investigate the utility of synthetic DoPaCos by comparing the performance of LED-large models without any further pre-training and those pre-trained with authentic vs. synthetic DoPaCos on the downstream note generation task. We then experiment with pre-training of the encoder and the decoder weights separately; we formulate "mix-and-match" experiments by combining encoder and decoder weights[4] from LED-large models pre-trained on only synthetic DoPaCos, only the clinical notes corpus, or both. Doing so results in nine possible combinations, and we proceed to assemble and fine-tune them on the same downstream training set. This experiment follows the intuition that synthetic conversations are more similar to the source format of the note generation task, while many documents from the clinical notes corpus highly resemble the destination format, i.e., HPI section of clinical notes, and it may be beneficial to expose the model to data of both modalities during pre-training.

To fine-tune the pre-trained models on the annotated DoPaCo dataset, we adopt an automated method for early stopping and selecting checkpoints for evaluation. At fixed intervals during fine-tuning, the geometric mean of ROUGE F1 scores is monitored, and training is stopped upon reaching the early-stopping patience. Then, the best checkpoint on the validation set is selected and evaluated on the test set. We do not monitor concept-based scores during training or validation, as doing so would result in excessively long training times. The hyper-parameters used for further pre-training and fine-tuning of LED models on all datasets are shown in Table 2. The average run-

[1] `https://huggingface.co/microsoft/DialoGPT-medium`

[2] Subjective, Objective, Assessment, and Plan

[3] `https://huggingface.co/allenai/led-large-16384`

[4] Decoder and LM head layers of each model are treated as decoder weights, with the rest of the parameters treated as encoder weights.

Table 2: *Hyper-parameters used for further pre-training and fine-tuning of the LED-large model.*

General			
Max. encoder length	5120	Max. decoder length	1024
Batch size	8	Max. gradient norm	1.0
Pre-training specific			
Learning rate	2.0×10^{-5}	Max. epochs	1
Window ratio	0.1	Max. window size	512
Warm-up ratio	0.01	Weight decay	0.001
Speaker mask ratio	0.5	Text infilling ratio	0.15
Fine-tuning specific			
Learning rate	0.001	Warm-up steps	200
Weight decay	0.001	Steps between evaluation	50
Early stopping patience	5	Max. generation length	512
Number of beams	5	No-repeat n-gram size	3

ning times for the pre-training and fine-tuning jobs on a single NVIDIA A10G GPU are 20 hours and 4 hours, respectively.

2.4. Evaluation metrics

We evaluate our systems using a mix of N-gram overlap-based, semantic similarity-based, and entity-based metrics. We employ ROUGE-$1/2/3/4/L$ [19] to compute N-gram (longest common sub-sequence for ROUGE-L) lexical overlaps between generated and human-written summaries.

To account for embedding-based semantic similarity [20], we use BERT score [21], which utilizes contextualized token embeddings obtained from a pre-trained Transformer model to compute the semantic similarity between generated and reference summaries. We use the official BERTscore package with the recommended DeBERTa-large-mnli model and rescale the scores with baselines. Both BERT scores and ROUGE$-3/4/L$ scores are found to correlate with human evaluation results in clinical note generation tasks [22].

To further investigate the conceptual relevance between generated and reference summaries, we rely on the Unified Medical Language System (UMLS) repository [23]. We extract the UMLS concepts from our generated summaries and compare them with those obtained from references using string matching algorithms; see, e.g., QuickUMLS [24]. By extracting and matching concepts, we can compute F1 scores and aggregate them across all test set conversations.

The extraction of biomedical concepts is based on string matching algorithms, which sometimes can lead to the extraction of irrelevant strings from the text. To circumvent this challenge, we leverage the clinical NER-based extraction method introduced in [25]. Specifically, we adopt a RoBERTa model [26] pre-trained on MIMIC-III clinical notes [27] and fine-tune it on the i2b2 2010 dataset [28] to extract clinical concepts. With this RoBERTa-based NER model, we can then match extracted concepts from generated and reference summaries to UMLS concepts and drop the ones we cannot find matches for. Finally, we can compute F1 scores across extracted concepts from the generated summaries and compare them with those from human-written summaries.

3. Results

In this section, we discuss the effects of in-domain pre-training using synthetic conversations. We first compare the performance of LED-large models without any further pre-training with those pre-trained with authentic or synthetic conversations on the DoPaCo summarization task in section 3.1. Then, the effects of pre-training with multiple data modalities are discussed in section 3.2, where the LED models are pre-trained on a combination of synthetic conversations and authentic clinical notes.

3.1. Pre-training using authentic vs. synthetic DoPaCos

Table 3 compares the performance of LED-large models pre-trained on authentic and synthetic DoPaCos to a baseline model without in-domain pre-training. The results show that ROUGE F1 scores and BERT scores consistently improve after further pre-training, regardless of whether authentic or synthetic conversations are used. This indicates that additional pre-training, even when using synthetic DoPaCos, may lead to improved overlap between the generated and reference summaries. This finding is consistent with the results of previous research [9].

In terms of UMLS and NER clinical concept-based metrics, however, pre-training with authentic DoPaCos did not significantly improve the baseline performance. This may be due to a stronger baseline compared to [9]. An edited set of training reference summaries with some boilerplate language removed was used for fine-tuning of all models in this work, and likely contributed to more medical concepts being included in the generated summaries. With synthetic data, we observed that concept extraction scores have decreased following pre-training. For the UMLS concept metrics, the baseline model achieved precision of 32.87% and recall of 51.12%, versus 31.10% and 50.08% from the model pre-trained on synthetic conversations. This may be due to that synthetic data is not as representative of real-world medical conversations as authentic data.

Overall, we find that further pre-training with authentic DoPaCos leads to improvements in terms of ROUGE and BERT scores, but the benefits on clinical concept-based metrics are relatively muted. A similar trend has been observed for synthetic data, but with slight decrease in performance across all metrics. Still, the model pre-trained with synthetic data outperforms the baseline model in 6 of 8 metrics. We conclude that although in-domain pre-training on synthetic data may not provide benefits surpassing pre-training on authentic data, it is indeed better than not performing any in-domain pre-training at all.

Apart from quantitative results as discussed above, qualitative human evaluations suggest that pre-training with conversations has led to perceivable improvements in the overall quality of the generated summaries. A comparison of sample generated summaries can be found in Table 4. We have also observed that further pre-training using DoPaCos resulted in improved stability (in terms of validation F1 scores, median generation length, and other monitored metrics) during fine-tuning.

3.2. Pre-training on multiple datasets

In Table 5 we present the test set results for the combination models. In most cases, combining pre-trained weights from two models harms the performance when compared to pre-training the encoder and decoder together. The exception is experiment No. 7, where the encoder pre-trained first on synthetic conversations and then on clinical notes (DoPaCos→Notes) is combined with a decoder pre-trained on synthetic DoPaCos only (DoPaCos), which outperforms many other combinations especially in ROUGE F1 scores.

We find that pre-training LED-large on only clinical notes (Experiment No. 5) leads to the worst performance among the experimented options. This could imply that pre-training on conversation-like data before fine-tuning on the conversation

Table 3: *Test set F1 scores (in percentages) of the LED-large models fine-tuned on the annotated DoPaCo dataset.*

Pre-training	**ROUGE Scores**					**Concept Scores**		**BERT Score**	**Median Length**
	R-1	R-2	R-3	R-4	R-L	UMLS	Clinical NER		
None	37.99	14.70	6.25	2.96	25.81	35.91	**49.84**	32.67	106.5
Authentic DoPaCos	38.21	14.79	**6.36**	3.10	26.12	**36.76**	49.19	32.94	105
Synthetic DoPaCos	**38.88**	**14.85**	6.35	**3.12**	**27.21**	34.20	48.39	**33.38**	97.5

Table 4: *An example DoPaCo with its reference (human-written) and generated HPI summaries. All generated summaries use wrong pronouns, since they were presented with de-identified conversation transcripts. The model with no DoPaCo pretraining is missing an important piece of information that "the patient's viral load is down to 90 copies". The italicized sentences are not factual.*

Summaries	**Conversation snippets**
Reference. The patient presents to the clinic for evaluation of HIV and HCV. She has completed a course of Harvoni. Her last HIV viral load was 90. She states she is feeling well.	[DR]: So, um, you took all the Harvoni, you took all the - [PT]: That is correct. [DR]: And your viral load is down to, to 90 copies. And the thing to tell you too is at some point the thing you're taking is going to be a single pill. [PT]: Okay. [DR]: So, that'll make life easier for you too, but your hep C, once that's treated, it usually does not come back. ... [DR]: How is the new cocktail? Any problems? [PT]: I was having them, um, every once in a while. I think - [DR]: Just as long as you eat with that thing, you'll be okay. [PT]: Yeah. Every once in a while, I'll have a, I call it a breakout. [LAUGHING]. [DR]: It's basically the toughest thing you can take for HIV... [DR]: We don't have the flu shot in our office this year, so I want you to run to like a CVS or a Walgreens - [PT]: No, I can't -... [DR]: Yeah. So, I'm pretty happy with this. It'll still Prezcobix. I think you're doing well on it, and you just keep going on. ... [DR]: So. You're doing well. I want to see you in like four months.
No DoPaCo pre-training. The patient presents to the clinic today for a follow-up visit. The patient is a **year old male presenting for f/u appointment. He is currently taking Harvoni and Prezcobix. He reports that he is compliant with his medications. He notes that he has been having occasional breakout of his HIV medication. He states that he does not have the flu shot.	
Pre-training with authentic DoPaCos. The patient is a male presenting to the clinic today for a follow-up visit. He is doing well overall. He has been compliant with all of his medications. The patient reports that his last viral load was 90 copies. He reports that he has not had any issues with his medication. *He does report that he does have occasional pruritus.* He denies any known side effects or adverse reactions.	
Pre-training with synthetic DoPaCos. The patient is a male presenting to the clinic today for a follow-up visit. Hepatitis C - He is currently taking Harvoni and is compliant with all of his medications. HIV - His last viral load was 90 copies. He is on Prezcobix.	

Table 5: *Test set F1 scores (in percentages) of the LED-large models pre-trained using various combinations of the two datasets: 160k synthetic conversations and 475k clinical notes. Best results for each metric are shown in bold.*

Index	**Pre-training**		**ROUGE Scores**					**Concept Scores**		**BERT Score**	**Median Length**
	Encoder	Decoder	R-1	R-2	R-3	R-4	R-L	UMLS	Clinical NER		
0	None		37.99	14.70	6.25	2.96	25.81	35.91	49.84	32.67	106.5
1	DoPaCos	DoPaCos	38.88	**14.85**	6.35	**3.12**	27.21	34.20	48.39	33.38	97.5
2	DoPaCos	Notes	37.56	14.49	6.10	2.86	25.61	36.59	47.80	32.23	97
3	DoPaCos	DoPaCos→Notes	38.55	14.59	6.25	3.01	27.09	32.42	47.35	32.36	92
4	Notes	DoPaCos	38.51	14.72	6.16	2.88	26.46	36.42	**52.07**	**33.81**	98
5	Notes	Notes	36.89	13.59	5.51	2.56	24.91	**37.57**	47.44	31.80	95
6	Notes	DoPaCos→Notes	38.34	14.72	6.15	2.88	26.66	35.09	46.68	32.90	86
7	DoPaCos→Notes	DoPaCos	**39.34**	14.77	**6.39**	**3.12**	**27.32**	35.58	51.83	33.38	98
8	DoPaCos→Notes	Notes	37.37	13.79	5.64	2.70	25.19	35.41	46.90	32.25	92
9	DoPaCos→Notes	DoPaCos→Notes	37.38	14.10	6.01	2.89	26.36	32.47	41.84	32.40	80

summarization task is beneficial, even if the data used for pre-training, in this case, are relatively noisy and of lower quality.

Contrary to our expectations, combining encoder weights pre-trained on conversations and decoder weights pre-trained on clinical notes (experiment No. 2) leads to low performance among the nine combinations. In fact, all combinations with decoders pre-trained only on clinical notes tend to underperform, especially in ROUGE and BERT metrics. Additionally, although pre-training the entire model on both datasets generally leads to high performance, the strongest model was obtained by combining encoder weights pre-trained on both datasets, and decoder weights pre-trained only on conversation data (Experiment No. 7). The underlying mechanism for this phenomenon is not entirely clear and warrants further investigation.

4. Conclusions

We studied pre-training practices for seq-to-seq models in automatic generation of clinical notes from doctor-patient conversations via abstractive summarization. A method for generating synthetic conversation snippets in medical domain using a dialog response generation model was described, and the synthetic conversations were used for pre-training of a summarization model. We found that pre-training with synthetic data was similarly beneficial as pre-training with authentic data, which suggests that authentic data could be saved for the fine-tuning stage. We also found that separated pre-training of the encoder and decoder weights in an encoder-decoder summarizer model using datasets of different modalities may be beneficial. This approach is inexpensive and viable when multiple datasets of distinctive nature are available for pre-training.

5. References

[1] N. L. Downing, D. W. Bates, and C. A. Longhurst, “Physician burnout in the electronic health record era: Are we ignoring the real cause?” *Annals of Internal Medicine*, vol. 169, no. 1, pp. 50–51, 2018, pMID: 29801050. [Online]. Available: https://www.acpjournals.org/doi/abs/10.7326/M18-0139

[2] A. D. Misra-Hebert, L. Amah, A. Rabovsky, S. Morrison, M. Cantave, C. A. Sinsky, and M. B. Rothberg, “Medical scribes: how do their notes stack up?” *Journal of Family Practice*, vol. 65, no. 3, pp. 155–160, 2016.

[3] R. Gidwani, C. Nguyen, A. Kofoed, C. Carragee, T. Rydel, I. Nelligan, A. Sattler, M. Mahoney, and S. Lin, “Impact of scribes on physician satisfaction, patient satisfaction, and charting efficiency: a randomized controlled trial,” *The Annals of Family Medicine*, vol. 15, no. 5, pp. 427–433, 2017.

[4] K. Krishna, S. Khosla, J. Bigham, and Z. C. Lipton, “Generating SOAP notes from doctor-patient conversations using modular summarization techniques,” in *Proc. of the 59th Annual Meeting of the ACL and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: ACL, Aug. 2021, pp. 4958–4972. [Online]. Available: https://aclanthology.org/2021.acl-long.384

[5] J. Su, L. Zhang, H. R. Hassanzadeh, and T. Schaaf, “Extract and abstract with BART for clinical notes from doctor-patient conversations,” in *Proc. INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, Incheon, Korea, Sep. 2022, pp. 2488–2492.

[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,” *J. Mach. Learn. Res.*, vol. 21, no. 1, Jan. 2020.

[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,” in *Proc. of the 58th Annual Meeting of the ACL*. Online: ACL, Jul. 2020, pp. 7871–7880. [Online]. Available: https://aclanthology.org/2020.acl-main.703

[8] L. Zhang, R. Negrinho, A. Ghosh, V. Jagannathan, H. R. Hassanzadeh, T. Schaaf, and M. R. Gormley, “Leveraging pretrained models for automatic summarization of doctor-patient conversations,” in *Findings of the ACL: EMNLP 2021*. Punta Cana, Dominican Republic: ACL, Nov. 2021, pp. 3693–3712. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.313

[9] C. Grambow, L. Zhang, and T. Schaaf, “In-domain pre-training improves clinical note generation from doctor-patient conversations,” in *Proc. of the First Workshop on Natural Language Generation in Healthcare*. Waterville, Maine: ACL, Jul. 2022, pp. 9–22. [Online]. Available: https://aclanthology.org/2022.nlg4health-1.2

[10] A. Yalunin, D. Umerenkov, and V. Kokh, “Abstractive summarization of hospitalisation histories with transformer networks,” *arXiv preprint arXiv:2204.02208*, 2022.

[11] Y. Yang, C. Malaviya, J. Fernandez, S. Swayamdipta, R. Le Bras, J.-P. Wang, C. Bhagavatula, Y. Choi, and D. Downey, “Generative data augmentation for commonsense reasoning,” in *Findings of the ACL: EMNLP 2020*. Online: ACL, Nov. 2020, pp. 1008–1025. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.90

[12] V. Kumar, A. Choudhary, and E. Cho, “Data augmentation using pre-trained transformer models,” in *Proc. of the 2nd Workshop on Life-long Learning for Spoken Language Systems*. Suzhou, China: ACL, Dec. 2020, pp. 18–26. [Online]. Available: https://aclanthology.org/2020.lifelongnlp-1.3

[13] B. Chintagunta, N. Katariya, X. Amatriain, and A. Kannan, “Medically aware GPT-3 as a data generator for medical dialogue summarization,” in *Proc. of the Second Workshop on Natural Language Processing for Medical Conversations*. Online: ACL, Jun. 2021, pp. 66–76. [Online]. Available: https://aclanthology.org/2021.nlpmc-1.9

[14] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, “DIALOGPT: Large-scale generative pre-training for conversational response generation,” in *Proc. of the 58th Annual Meeting of the ACL: System Demonstrations*. Online: ACL, Jul. 2020, pp. 270–278. [Online]. Available: https://aclanthology.org/2020.acl-demos.30

[15] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document transformer,” *arXiv preprint arXiv:2004.05150*, 2020.

[16] C. Meister, T. Pimentel, G. Wiher, and R. Cotterell, “Locally Typical Sampling,” *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 102–121, 01 2023. [Online]. Available: https://doi.org/10.1162/tacl_a_00536

[17] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, “Transformers: State-of-the-art natural language processing,” in *Proc. of the 2020 EMNLP: System Demonstrations*. Online: ACL, Oct. 2020, pp. 38–45. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.6

[18] M. Zhong, Y. Liu, Y. Xu, C. Zhu, and M. Zeng, “Dialoglm: Pre-trained model for long dialogue understanding and summarization,” in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11 765–11 773.

[19] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out*. Barcelona, Spain: ACL, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[20] W. Kryscinski, B. McCann, C. Xiong, and R. Socher, “Evaluating the factual consistency of abstractive text summarization,” in *Proc. of the 2020 EMNLP (EMNLP)*. Online: ACL, Nov. 2020, pp. 9332–9346. [Online]. Available: https://aclanthology.org/2020.emnlp-main.750

[21] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with bert,” in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[22] F. Moramarco, A. Papadopoulos Korfiatis, M. Perera, D. Juric, J. Flann, E. Reiter, A. Belz, and A. Savkov, “Human evaluation and correlation with automatic metrics in consultation note generation,” in *Proc. of the 60th Annual Meeting of the ACL (Volume 1: Long Papers)*. Dublin, Ireland: ACL, May 2022, pp. 5739–5754. [Online]. Available: https://aclanthology.org/2022.acl-long.394

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

[24] L. Soldaini and N. Goharian, “Quickumls: a fast, unsupervised approach for medical concept extraction,” in *Proc. of the 2nd SIGIR workshop on Medical Information Retrieval (MedIR)*, 2016, pp. 1–4.

[25] X. Yang, J. Bian, W. R. Hogan, and Y. Wu, “Clinical concept extraction using transformers,” *Journal of the American Medical Informatics Association*, vol. 27, no. 12, pp. 1935–1942, 2020.

[26] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.

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

[28] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, “2010 i2b2/va challenge on concepts, assertions, and relations in clinical text,” *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 552–556, 2011.