Clinical Note Dialogue Generation in the Medical Domain

Toma Sabin-Sebastian Popa Andrei-Ionut

Abstract

This report presents a study conducted for the MEDIQA-Chat shared task at ACL ClinicalNLP, focused on the bidirectional generation between clinical notes and doctor-patient dialogues. We explore various transformer-based models such as T5-small, T5-base, and FLAN-T5, and compare their performance with ChatGPT-4. Evaluation is conducted using standard metrics such as BLEU, ROUGE, and METEOR. The experiments aim to support data augmentation and automation in the medical domain.

1. Objective

The primary objective of this project is to develop and evaluate models capable of converting clinical notes to dialogues and vice versa. This supports medical data augmentation, automated documentation, and clinical NLP advancements.

2. Datasets

- **MTS-Dialog:** The MTS-Dialog dataset is a large-scale, multi-turn, multi-section dataset designed for the task of clinical note generation from doctor-patient dialogues. It consists of over 13,000 real-world conversations sourced from the MIMIC-IV database, each aligned with structured clinical notes divided into standard sections like Chief Complaint, History of Present Illness, Past Medical History, and more. Each dialogue reflects a natural, turn-by-turn exchange between a healthcare provider and a patient, capturing the nuances and variability of real clinical interactions.
- **ACI-BENCH:** The ACI-BENCH (Ambient Clinical Intelligence Benchmark) dataset is the largest publicly available corpus of doctor–patient visits paired with clinical notes, created to advance and evaluate automated note generation from clinical dialogues. ACI-BENCH supports various real-world note-taking scenarios: natural ambient conversations, dialogues containing scribe-assisted dictations, and interactions involving virtual assistant prompts. It also includes both human-transcribed audio and ASR-generated transcripts, allowing researchers to benchmark the impact of transcription quality on note generation

3. Subtask A – Dialogue2Note Generation

3.1 T5-small (Baseline)

- Trained on 100 samples, 25 for evaluation, 10 for testing, 1 epoch.
- Performance:
  - Low ROUGE-1 and ROUGE-L.
  - Validation ROUGE-1 is around 0.19, and ROUGE-L about 0.14. This indicates some overlap of unigrams and longest common subsequences with the reference notes, but relatively low overall.
  - Test performance drops slightly, especially in ROUGE-L, which shows weak structural coherence in the generated notes.
  - BLEU on the test set is 0.0, meaning no n-gram matches (likely up to 4-gram) between the model's output and reference notes — a red flag for fluency or relevance.
  - BLEU is generally harsher than ROUGE in low-resource abstractive summarization, but 0.0 suggests severely mismatched phrasing.
  - Model exhibits repetitive and verbatim outputs.
- Example predictions:
  1. **Doctor:** Do you know about allergies from any medications? **Patient:** No.
  2. **Doctor:** It is good to see you again. How have you been? **Patient:** It is good to see you too. I have been good. **Doctor:** Have any changes to your medical history or social history since last time I saw you? **Patient:** No. No changes.
  3. **Doctor:** Are you still working, sir? **Patient:** Are you still working, sir? **Patient:** No, I am retired now. I worked for twenty years in social security administration. **Patient:** Now I just enjoy my life with my wife, we travel a lot. I like to be extremely active. **Doctor:** How about smoking or any drugs? **Patient:** No, I don't smoke or do any kind of drugs or anything.
- Model is repeating or copying directly, not summarizing into a structured clinical note.

3.2 T5-small (Full Dataset)

- Trained for 3 epochs.
- Performance:
  - Validation ROUGE-1 is around 0.19, and ROUGE-L about 0.14. This indicates some overlap of unigrams and longest common subsequences with the reference notes, but relatively low overall.

  - Test performance drops slightly, especially in ROUGE-L, which shows weak structural coherence in the generated notes.
  - BLEU on test set is 0.0, meaning no n-gram matches (likely up to 4-gram) between your model's output and reference notes — this is a red flag for fluency or relevance.
  - BLEU is generally harsher than ROUGE in low-resource abstractive summarization, but 0.0 suggests severely mismatched phrasing.

- Example predictions:

  1. The patient has had a history of asthma.
  2. The patient has been checking blood sugar twice a day and records the readings each time. The patient has been experiencing any dizziness, excessive urination, fatigue, weight loss, or any other symptoms. He has a copy of his blood work right here and it looks like his cholesterol is still pretty high. He wants to continue taking the cholesterol medication I prescribed and monitoring his cholesterol intake. He needs to schedule another follow up appointment and get you in in a few weeks after another blood draw.
  3. The patient is in good shape. She has no heart issues, genital or urinary tract issues. She has no muscle or movement issues. She has no muscle or movement issues. She has no muscle or movement issues.

- The model shows basic understanding of medical dialogue but struggles with factual accuracy and abstraction. It often hallucinates details, repeats phrases, and copies input verbatim instead of generating structured clinical notes.

3.3 FLAN-T5 Base

- Trained for 2 epochs.
- Performance:
  - The model is beginning to capture relevant content from the dialogues more consistently.
  - ROUGE-1 and ROUGE-L are noticeably improved compared to the earlier runs, indicating better lexical and structural overlap with reference summaries.
  - ROUGE-L improvement suggests better handling of sequence structure and note format.
  - BLEU scores remain low, especially on the test set — suggesting limited n-gram precision and that the model may still struggle with grammatical fluency or verbatim phrasing.

Example predictions:

1. Patient has a history of blood clots.

2. Patient's father died of thoracic aortic aneurysm and his mother died of stroke.
3. Patient has been feeling better lately.
4. Patient is allergic to Bactrim and adhesive tape.
5. Patient has a migraine headache
6. Patient doesn't have a history of alcohol or drug abuse.
7. Patient is divorced. He works as a homemaker.

3.4 T5-base (Final Model)

- **Framework:** HuggingFace Trainer.
- Max Input Length: 512 tokens, Output: 128 tokens.
- Batch size: 2, Epochs: 3, Mixed Precision (FP16) enabled.
- Results: Moderate ROUGE and METEOR scores, low BLEU.
- The T5 model performs reasonably well in generating semantically relevant outputs, as shown by the ROUGE and METEOR scores. Although the BLEU score is low (common in abstractive tasks), the results suggest that the model captures key information and maintains partial structure from the reference dialogues.

4. Subtask B – Full Note Generation

- Objective: Generate comprehensive clinical notes from full doctor-patient conversations, covering all relevant sections (e.g. Past Medical History, Assessment).
- Model: Reused T5 trained on Dialogue2Note.
- The model generalizes moderately well to unseen data. ROUGE and METEOR scores indicate the model preserves semantic relevance and captures partial structure of full notes, though BLEU shows limitations in exact phrasing.
- Example predictions:
  - The patient is a 2-and-a-half-year-old male and a 5-and-a-half-year-old stepson who are sick at home. The little one was taken to ER with an inhaler of albuterol. The little one was sick with flu shortly after delivery and the mother was sick with flu shortly after delivery. The patient's mother was visiting dad's family. The family does smoke at home.

5. ChatGPT-4 Comparison – Subtask A and B

- ChatGPT-4 generates highly structured and semantically rich summaries from dialogues.
- Captures nuanced medical history and presents them fluently.
- Superior to T5 models in content relevance and coherence.
- Example predictions subtask A:
    1. Patient presents with congestion and frequent coughing, reports choking sensation.
    2. Lives with parents, 2.5 y/o brother, and 5.5 y/o stepbrother. Siblings have cold symptoms and vomiting. Recent travel to father's family. Exposure to indoor pets. Positive smoke exposure from mother.
    3. No known drug allergies.
    4. History of Huntington disease, dementia, and hypertension. Followed by PCP, neurologist (Dr. Townsend), and psychiatrist (Dr. Smith).
    5. Resides at Life Care Center. Denies tobacco, alcohol, or IV drug use.
    6. Spontaneous miscarriage at 3 months gestation, approximately 30 years ago.
    7. 83-year-old female with progressive dysphagia (1 year), dysarthria, leg cramps, and worsening bilateral arm weakness.
    8. Stable condition.
    9. Recent pneumonia, urosepsis, dementia, left toe amputation, osteoporosis, and hypothyroidism.
- Example predictions subtask B:
    1. The patient presents with complaints of significant congestion and persistent coughing. She reports a sensation of choking associated with the symptoms.
    2. The patient lives with her mother, father, a 2.5-year-old brother, and a 5.5-year-old maternal stepbrother, both of whom are currently ill with cold symptoms including vomiting and diarrhea. The younger brother was evaluated in the ER and discharged with an albuterol inhaler. A maternal nephew is also reported to have an ear infection. The patient's mother reports she had the flu shortly after delivery. There was recent holiday travel to visit the paternal side of the family. The household includes exposure to cats and dogs indoors. The patient has second-hand smoke exposure from the mother.
    3. The patient denies any known drug allergies.
    4. The patient reports no significant past medical history. Denies history of chronic diseases including hypertension, depression, or malignancy.
    5. An 83-year-old woman presenting with progressive dysphagia for the past year, dysarthria, right arm weakness initially, leg cramps, and now progressive bilateral upper extremity weakness.

6. Subtask C – Note2Dialogue Generation

Training Configuration Comparison

Setting	Model 1 (Default)	Model 2 (Tuned)
Epochs	3	10
Learning Rate	5e-4	3e-4
Batch Size	2	4
Warmup Steps	0	500
Gradient Accumulation Steps	1	2

Table 1: Training setup comparison between the default and tuned models.

Task Objective

The goal of this task is to generate a synthetic doctor-patient conversation from a given clinical note. These notes typically contain information such as patient history, symptoms, diagnosis, treatments, or recommendations. The model should reconstruct a natural and realistic dialogue that would logically produce the clinical note content.

Performance and Observations

- The tuned model outperformed the default model across all key metrics (ROUGE-1, ROUGE-L, METEOR), indicating improved coherence and semantic relevance in generated conversations.
- BLEU scores remained low for both models, which is expected in open-ended generation tasks where exact phrasing overlap is rare.
- The second (tuned) model benefited from a more stable configuration with a reduced learning rate, increased epochs, warmup steps, and gradient accumulation. These changes led to better generalization and robustness.
- Fine-tuning significantly improved the model's ability to generate meaningful and logically connected dialogue structures.

Example Prediction (Tuned Model)

```
    [doctor] hi, ms.  lee.  how are you?
patient
    i'm doing okay.
doctor
    i'm doing okay.
patient
```

```
     i'm doing okay.
doctor
     i'm doing okay.
patient
     i'm doing okay.
doctor
     i'm doing okay.
patient
     i'm doing okay.
doctor
     okay.
patient
     i was at a birthday party and i ran my back into the railing of the
     staircase and twisted my back.
doctor
     okay.
patient
     i ran my back into the railing of the staircase and twisted my back.
doctor
     okay.
patient
     i ran my back into the railing of the staircase and twisted my back.
doctor
     okay.
```

This example highlights the model's tendency to repeat phrases and fallback into loops despite showing the ability to extract incident-related information.

7. Conclusion

- We trained T5 models to generate clinical notes and synthetic dialogues from doctor-patient data. Results show promising coherence, especially for Dialogue2Note, with room for improvement through tuning and post-processing.
- BLEU is not a fully reliable metric for abstractive clinical summarization.
- ChatGPT-4 demonstrates superior fluency and contextual understanding.
- Fine-tuning and parameter adjustments are critical for improving generation quality.

Future Work

- Train on Transcripts from Real Audio Use ASR outputs (automatic speech recognition) to simulate real-world use cases.
- Evaluate with Medical Professionals Collect qualitative feedback from clinicians to assess realism, usefulness, and trustworthiness.
- Integrate Post-processing Heuristics Automatically correct hallucinated medical terms or ensure section consistency in generated dialogues.