Summarizing Diagnoses from Clinical Notes: Towards a Benchmark and Systematic Evaluation

Anonymous Authors

Summarizing diagnoses from clinical notes (long, unstructured text) can reduce the diagnostic documentation burden and support handoffs, referrals, and follow-ups. Towards this goal, our work introduces a reproducible benchmark and systematic evaluation for diagnosis summarization from clinical notes [1], with two input settings framing diagnosis summarization as a sequence-to-sequence generation task, with discharge diagnoses of notes serving as targets. To reflect real-world clinical variability, two input formulations are: ***Summ-CPP***, focusing on chief complaint, history of present illness, and past medical history; and ***Summ-Full-Note***, which incorporates the entire note, excluding diagnoses. Our study evaluates both general-purpose and domain-adapted language models under fine-tuned and zero-shot prompt configurations. Pretrained language models (BART, T5, and SciFive) are fine-tuned on curated training notes, while large language models (e.g., GPT-4o-mini) are evaluated in a zero-shot prompt setting. We will release a curated dataset on PhysioNet[1] under the MIMIC-IV-notes [1] data use agreement. Curated dataset consists of 10k samples, balanced across service types (medicine, surgery, orthopaedics, neurology, cardiothoracic, neurosurgery, obstetrics, psychiatry, urology, and plastic surgery) and divided into three splits: train (8k), validation (1k), and test (1k). Performance is assessed using ROUGE and BERTScore to quantify both n-gram overlap and semantic similarity, with results summarized in Table 1. Our results show that domain-adapted models are crucial for achieving clinically reliable performance.

Table 1: Summary of the percentage F1-scores (ROUGE and BERTScore) of all models across both input settings, where R1, R2, and RL represent ROUGE-1, ROUGE-2, and ROUGE-L, respectively.

Model	**Summ-CPP**				**Summ-Full-Note**			
	R1	R2	RL	BERTScore	R1	R2	RL	BERTScore
BART-base	36.83	21.95	35.88	87.50	33.87	20.05	33.90	84.67
BART-large	39.59	24.30	38.03	87.89	35.47	22.00	35.84	81.20
T5-base	53.61	38.30	52.88	92.16	54.21	38.92	53.49	92.16
T5-large	53.57	37.94	52.82	91.86	55.52	40.01	54.58	92.20
SciFive-base	55.10	40.20	54.30	91.20	56.42	40.86	55.00	91.50
SciFive-large	56.29	41.21	55.52	92.46	57.00	42.00	56.20	92.62
LLaMa-3-8B-Instruct	52.80	37.90	52.10	89.40	53.50	38.20	52.40	89.70
BioMistral-7B-DARE	51.50	36.50	51.00	87.80	52.00	37.32	51.46	88.28
GPT-4o-mini	54.20	39.24	53.53	90.40	55.14	40.00	54.86	90.60

References

[1] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

[1] https://physionet.org