Large Scale Sequence-to-Sequence Models for Clinical Note Generation from Patient-Doctor Conversations

Gagandeep Singh, Yue Pan, Jesús Andrés Ferrer, Miguel Del-Agua Teba
Frank Diehl, Joel Pinto, Paul Vozila
Nuance Communications
1 Wayside Road, Burlington, MA 01803
{gagandeep.singh1, yue.pan, jesusandres.ferrer, miguel.delagua,
frank.diehl, joel.pinto, paul.vozila}@nuance.com

Abstract

We present our work on building large scale sequence-to-sequence models for generating clinical note from patient-doctor conversation. This is formulated as an abstractive summarization task for which we use encoder-decoder transformer model with pointer-generator. We discuss various modeling enhancements to this baseline model which include using subword and multiword tokenization scheme, prefixing the targets with a chain-of-clinical-facts, and training with contrastive loss that is defined over various candidate summaries. We also use flash attention during training and query chunked attention during inference to be able to process long input and output sequences and to improve computational efficiency. Experiments are conducted on a proprietary dataset containing about 900K encounters in U.S. English from around 1800 healthcare providers covering 27 specialties. The results are broken down into primary care and non-primary care specialties. Consistent accuracy improvements are observed across both of these categories.

1 Introduction

Medical documentation plays an important role in diagnosis, treatment, and delivery of safe patient care. Healthcare professionals are required by law to document their encounter with the patient. Apart from this, medical documentation is also useful in research and driving quality improvement (Menachemi and Collum, 2011). Medical documentation needs to be accurate and comprehensive, capturing the patient history, physical examination, laboratory and imaging studies, diagnosis, and treatment options. Physicians typically spend 35% of their time on documenting the patients visits of the day (Joukes et al., 2018). This increased documentation burden is one of the main causes for physician burnout (Wright and Katz, 2018; van Buchem et al., 2021).

The use of automatic speech recognition (ASR) systems have simplified document creation to a great extent where physicians dictate medical notes into electronic health records (EHRs). The content in the dictation is by and large already discussed with the patient albeit in colloquial language. Advances in deep learning in the field of natural language processing has attracted increased attention in generating medical reports directly from patient-doctor conversations (Krishna et al., 2021; Enarvi et al., 2020; Joshi et al., 2020; Michalopoulos et al., 2022; Zhang et al., 2021). Some of the challenges posed by this problem are:

1. The transcripts can be long, reaching 10k words for a 53 minute patient encounter (including punctuation and special tokens). This poses modeling challenges as well as computational challenges.

2. The conversational nature of interaction with long range context is difficult to summarize compared to one contiguous stretch of transcript or document.

3. The transcript language is very informal compared to medical reports, with usage of colloquial terminology, e.g., belly for abdomen, and might have incomplete information that was conveyed visually, e.g., a patient might point and say "it hurts here".

Encouraged by ongoing advancements in neural sequence transduction (e.g., for machine translation and abstractive summarization), we follow an end-to-end approach to the problem. We use a single transformer model to generate clinical reports directly from patient-doctor conversational transcripts with various enhancements to handle long input and output sequences. Our approach is similar to Enarvi et al. 2020 where a transformer model with pointer generator was used to generate clinical notes for Orthopedics. We extend this with

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 138–143

Partition	Primary Care	Non-Primary C.
train	489k	372k
recent	70k	53k
dev	2.4k	2.6k
test	21k	15k

Table 1: Number of encounters breakdown

various modeling improvements that are discussed in Section 3.

2 Dataset

We use a dataset consisting of medical encounters across 27 medical specialties in the ambulatory setting. Each encounter includes a patient-doctor conversation transcribed and diarized by an automatic speech recognition (ASR) system. The ASR transcript is used to generate three sections of a medical note, namely History of Present Illness (HPI), Assessment and Plan (AP), and Physical Examination (PE). The median number of words in each of these sections is 166, 291, and 111 respectively; while that for the transcript is 2128. The dataset is collected across 128 medical institutions and 1811 physicians.

3 Modeling

We use a sequence-to-sequence model with transformer architecture (Vaswani et al., 2017) and train a separate model for each of the three sections. Since the report format and style varies across specialties and physicians, each transcript is prepended with a unique specialty ID and doctor ID to condition the report generation. In all of our experiments we use the big model size, similar to the one specified in Vaswani et al. 2017 with 16 attention heads in each multi-head attention module, inner representations of size 1024, and the feed-forward layer size of 4096 in each transformer layer. We, however, use an 8 encoder layers and 4 decoder layers configuration instead of the default 6-6 one since the transcripts are longer and have a higher perplexity language than the reports. We use pre-layer normalization (Baevski and Auli, 2019) and the pointing mechanism (See et al., 2017). For positional encoding, on encoder side we use rotary positional embeddings (RoPE) (Su et al., 2021) and on decoder side we use the T5 scalar relative positional embeddings (Raffel et al., 2020). We make several changes over this baseline model in order to further tailor it to our problem as discussed in the following subsections.

3.1 Modeling Enhancements

3.1.1 Subword and Multiword Tokenization

Word based vocabulary systems replace any word outside of the fixed vocabulary with an out of vocabulary `OOV` token. Most language generation systems use subword modeling to create an open vocabulary system (Sennrich et al., 2016; Schuster and Nakajima, 2012; Kudo and Richardson, 2018). Subword modeling alone increases sequence length versus a word-based encoding, exacerbating the challenge of handling very long medical conversations.

Additionally, medical reports often contain templates[1] that occur very frequently, suggesting subsequences may be modeled atomically. To support an open vocabulary without compromising sequence length, we used SentencePiece (Kudo and Richardson, 2018) and specified 'space' as a regular character so that word boundaries do not enforce token boundaries. Training a SentencePiece model in such a manner leads to an open vocabulary system that includes subwords as well as multiwords.

3.1.2 Chain-of-Clinical-Facts

In order to help the model learn an intermediate summary plan while doing abstractive summarization, Narayan et al. 2021 proposed prepending target summaries with an ordered sequence of entities mentioned in the summary. Motivated by this, we trained the model to generate a chain-of-clinical-facts that are present in the summary before generating the summary. These facts were extracted from the reference summaries using a proprietary fact extraction tool that tags the clinically relevant words in the summary. Examples include the words that convey symptoms, diagnosis, treatment, etc., along with qualifying attributes e.g., body part, laterality, severity, etc. Thus the decoder first generates an executive summary of the medical note before generating the full medical note, and consequently the generated medical note is conditioned both on the transcript as well as the relevant medical facts. During inference, no external fact extraction is needed and the generated chain-of-facts can be discarded.

[1] designed as typing/dictation accelerant and for increasing note consistency

Section	Model	Primary Care	Non-Primary Care
AP	Baseline	62.9 / 62.5 / 50.4	68.1 / 67.6 / 56.3
	+ subword & multiword	64.9 / 64.1 / 52.0	69.5 / 70.0 / 59.4
	+ chain-of-facts	65.3 / 65.2 / 53.3	70.0 / 70.6 / 59.9
	+ contrastive loss	**66.2 / 65.7 / 53.6**	**70.9 / 71.2 / 60.3**
HPI	Baseline	44.5 / 60.9 / 42.5	49.3 / 62.5 / 45.6
	+ subword & multiword	46.2 / 61.0 / 42.8	51.2 / 63.2 / 46.7
	+ chain-of-facts	46.5 / 61.6 / 43.4	51.1 / 64.1 / 47.4
	+ contrastive loss	**47.7 / 62.3 / 43.9**	**52.5 / 64.7 / 47.9**
PE	Baseline	78.2 / 77.6 / 74.8	80.8 / 81.2 / 77.8
	+ subword & multiword	**80.0 / 79.5 / 77.0**	**82.4 / 83.2 / 79.5**

Table 2: Accuracy with various modeling techniques; the three F1 scores per cell are: ROUGE-L / Fact-C / Fact-F

3.1.3 Contrastive Loss

During training we applied the BRIO constrastive loss introduced in Liu et al. 2022 to enhance the accuracy of probability estimation for system-generated summaries, rather than relying solely on teacher-forced cross-entropy training. This contrastive loss is defined by

$$L_{ctr} = \sum_{i=0}^{K-1} \sum_{j>i}^{K-1} max(0, f(S_j) - f(S_i) + \lambda_{i,j}) \quad (1)$$

where S_i and S_j are two out of K candidate summaries and $SCORE(S_i) > SCORE(S_j), \forall i < j$. $\lambda_{i,j}$ is the ranking margin between the two candidates as in the original BRIO paper. $f(S_i)$ is the length-normalized estimated log-probability. This produces $\binom{K}{2}$ comparisons for each encounter.

In general, better results can be achieved by using a larger number of candidates with GPU memory being the bottleneck. To address this issue, we implement a strategy where $K-1$ out of N candidates are randomly sampled for each encounter within a batch while always keeping the top ranked hypothesis. The N candidates are generated by the cross-entropy trained baseline model using nucleus sampling (Holtzman et al., 2020). During training, we combine the contrastive and cross-entropy loss to use the model trained directly to generate the summary, instead of having to re-rank candidates generated by the cross-entropy trained model.

3.2 Speed and Memory Efficiency Enhancements

Due to the long input sequences, we adopted Flash Attention (Dao et al., 2022) for encoder self-attention during training which provided large memory savings and training speed-up. We explored using it for decoder self-attention and encoder-decoder cross-attention as well, but the incremental efficiency gain was limited. During inference, in order to compute full attention in a memory-efficient manner across a wide range of GPUs without requiring corresponding Flash Attention kernels, we process self-attention queries in chunks, as suggested in Gupta et al. 2021.

4 Evaluation metrics

We report three F1 score-based accuracy metrics: (a) ROUGE-L: This is our implementation of the ROUGE metric (Lin, 2004) in which we check for the longest common subsequence between the reference and hypothesis; (b) Fact-C: This measures the overlap of core medical facts, e.g., pain, automatically extracted from the hypothesis and reference; (c) Fact-F: This reflects the match of full medical facts, including attributes, e.g., laterality, body part.

5 Experiments

We trained our models on 4 $\times$ 80GB GPU machines with data-parallel training using the fairseq library (Ott et al., 2019). Each model was trained for a predefined number of steps on the train partition, and decoded and scored at multiple checkpoint intervals. The test partition contains the chronologically latest encounters for each physician, while dev contains the set of encounters just before test for each physician. We also create a smaller subset of the train partition called recent that consists of the latest 200 encounters for each physician. It is used to fine-tune the trained model to the most recent encounters so as to bias it to-

Tokenization	max train src / tgt tokens len	Train Steps	Accuracy
word-vocab	4096 / 1536	20k	44.5 / 60.9 / 42.5
word-vocab*	6144 / 1856	30k	44.4 / 60.3 / 42.0
subword SPM	4096 / 1536	20k	45.2 / 61.3 / 43.0
subword SPM*	6528 / 1962	32.5k	45.4 / 61.2 / 43.2
sub/multi-word*	4096 / 1536	20k	46.5 / 61.0 / 42.8

Table 3: Comparison of various tokenization techniques. Accuracy is reported on HPI section of primary care and the reporting format is same as in Table 2. * correspond to experiments that used the same effective average input and target length in terms of number of words and were trained to the same number of epochs

wards evolving style, templates, etc. The number of encounters breakdown per partition is shown in Table 1. Specialties with fewer encounters were sampled more often during training. We averaged the last 10 model checkpoint weights to reduce the variance in results and picked the best performing averaged checkpoint on the dev set to report test results. For all experiments we use a vocabulary size of 45k tokens which is shared for the source and the target. Encoder and decoder token embeddings are also shared. We calculate and report micro averages which are broken down into 1) primary care specialities, which consist of family medicine and internal medicine, and 2) all other specialties that we refer to as non-primary care. The primary care specialties deal with a broad set of diseases and conditions for people of all ages and are thus harder to model.

For tokenization experiments, we trained the SentencePiece model on the train partition. Token length was restricted to 100 characters. We also reserved certain words to be included in the vocabulary, such as the specialty IDs, patient IDs and speaker turn indicators.

For chain-of-clinical-facts experiments, we prepended the facts to the summary with a <SEPARATOR> token in between, with individual fact phrases separated by a <FACT_SEP> token. On an average, the length of the prefix is about 20% that of reports, excluding the separator tokens.

For the contrastive loss training process, we generated 20 candidate summaries for each encounter in the recent partition using the base model that was trained with cross-entropy loss. We applied nucleus sampling with a probability mass of 0.6 to generate these summaries. We then ranked the summaries based on their average ROUGE-L and Fact-C scores, with the highest-scoring summary being ranked first. Finally, we fine-tuned the base model using an equal-weighted combination of the contrastive and cross-entropy loss. During fine-tuning, we dynamically chose 8 out of the 20 candidates for each example in the batch for computation and memory efficiency, where the top ranked hypothesis was always kept, while the rest 7 were sampled randomly.

6 Results

The accuracy for each of the three sections by incrementally adding the various modeling technique is shown in Table 2. The baseline is a transformer pointer-generator. There is a general trend of improvement over all categories as the proposed model components are added. We did not observe any improvement to the physical exam (PE) section from the use of chain-of-clinical facts and contrastive loss which is probably due to the heavily templated nature of documentation in this section.

The use of subword and multiword tokenization, apart from improving accuracy, also helps to speed up model convergence as seen in Table 3. Due to the nature of subword & multiword tokenization, the system benefited from 1) more number of epochs for the same number of training steps; 2) longer training context at the same effective length in terms of number of words.

With the use of Flash Attention, we were able to increase the number of tokens per batch by 4x yielding a training speed-up of 2-2.5x times in terms of number of tokens processed per second. During inference, query chunked attention enables processing transcripts of any length without truncation as opposed to vanilla attention which runs out of memory on a 16G GPU for inputs longer than 10k tokens.

7 Conclusions

We used transformer-based models to build a large-scale, multi-specialty, end-to-end abstractive summarization system capable of generating medical

reports from conversations. We presented various modeling and efficiency improvements that can be applied to better adapt these models to this challenging task.

References

Alexei Baevski and Michael Auli. 2019. Adaptive input representations for neural language modeling. In *International Conference on Learning Representations*.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy. 2020. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 22–30, Online. Association for Computational Linguistics.

Ankit Gupta, Guy Dar, Shaya Goodman, David Ciprut, and Jonathan Berant. 2021. Memory-efficient transformers via top-k attention. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 39–52, Virtual. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Online. Association for Computational Linguistics.

Erik Joukes, Ameen Abu-Hanna, Ronald Cornet, and Nicolette F de Keizer. 2018. Time spent on dedicated patient care and documentation tasks before and after the introduction of a structured and standardized electronic health record. *Applied clinical informatics*, 9(01):046–053.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Nir Menachemi and Taleah H Collum. 2011. Benefits and drawbacks of electronic health record systems. *Risk management and healthcare policy*, 4:47.

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. 2022. MedicalSum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

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, Vancouver, Canada. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *CoRR*, abs/2104.09864.

Marieke M van Buchem, Hileen Boosman, Martijn P Bauer, Ilse MJ Kant, Simone A Cammel, and Ewout W Steyerberg. 2021. The digital scribe in clinical practice: a scoping review and research agenda. *NPJ digital medicine*, 4(1):57.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alexi A. Wright and Ingrid T. Katz. 2018. Beyond burnout — redesigning care to restore meaning and sanity for physicians. *New England Journal of Medicine*, 378(4):309–311. PMID: 29365301.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R. Gormley. 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3693–3712, Punta Cana, Dominican Republic. Association for Computational Linguistics.