DeakinNLP at ProbSum 2023: Clinical Progress Note Summarization with Rules and Language Models

Ming Liu[1], **Dan Zhang**[1], **Weicong Tan**[2], **He Zhang**[3]
[1] School of IT, Deakin University, Australia
[2] Faculty of IT, Monash University, Australia
[3] Zhongtukexin Co. Ltd. , China
{m.liu, dan.zhang}@deakin.edu.au
weicong.tan@monash.edu
zhanghe@kxsz.net

Abstract

This paper summarizes two approaches developed for BioNLP2023 workshop task 1A [1]: clinical progress note summarization. We develop two types of methods with either rules or pre-trained language models. In the rule-based summarization model, we leverage UMLS (Unified Medical Language System) and a negation detector to extract text spans to represent the summary. We also fine tune three pre-trained language models (BART, T5 and GPT2) to generate the summaries. Experiment results show the rule based system returns extractive summaries but lower ROUGE-L score (0.043), while the fine tuned T5 returns a higher ROUGE-L score (0.208).

1 Introduction

Clinical progress note is a typical type of text format for doctors and nurses to keep a record of a patient's up-to-date status. Automatic summarization of daily progress notes can not only saves doctors' time, but also benefit the structured representation for patient record. In BioNLP workshop task 1A [2], the task of problem list summarization (Gao et al., 2023) aims to generate a list of diagnoses and problems in a patient's daily care plan using input from the providers' progress notes during hospitalization. Different from generic text summarization, the amount of the note and summary pairs is limited, e.g., there are only 765 annotated records which can be used for training, compared to the hundreds of thousands in XSum (Narayan et al., 2018). Recent methods (Gao et al., 2022) used pre-trained language models, such as T5 (Raffel et al., 2020) and BART (Lewis et al., 2019), to adapt to the task of clinical text summarization.

In this paper, we will show the two basic methods that we used for this summarization task: The first one is a rule based system which leverages UMLS and a negation detector for high frequency medical concept filtering, and these concepts are used to extract text spans. The second approach fine tunes pre-trained language models, including BART (Lewis et al., 2019), T5 (Raffel et al., 2020) and GPT2 (Radford et al., 2019). Experiments on a blind test set shows that the fine tuned T5 achieves the best performance, the result is well aligned with that in (Gao et al., 2022).

2 Related Work

Recent work on clinical text summarization focus on two applications: pure clinical note summarization (Liang et al., 2019; Adams et al., 2021) and clinical dialogue summarization (Yim and Yetisgen-Yildiz, 2021; Zhang et al., 2021). Liang et al. (2019) proposed three sentence classification models that extracts sentences from progress notes. Zhang et al. (2021) constructed an English dataset of 109,000 hospitalizations (2M source notes) and their corresponding summary proxy. More recently, Kanwal and Rizzo (2022) used multi-head attention-based mechanism to perform extractive summarization of meaningful phrases on clinical notes. Chuang et al. (2023) introduced a soft prompt-based calibration on mitigating performance variability in clinical Notes summarization.

3 Exploration of Clinical Progress Notes

The clinical progress notes were extracted from MIMIC-III, a publicly available dataset of de-identified EHR data from approximately 60,000 hospital ICU admissions at Beth Israel Deaconess Medical Center in Boston, Massachusetts (Johnson et al., 2016). A subset of the above progress notes were annotated with the SOAP format with four components: *Subjective*, *Objective*, *Assessment*,

[1] https://physionet.org/content/bionlp-workshop-2023-task-1a/1.0.0/
[2] https://physionet.org/content/bionlp-workshop-2023-task-1a/1.0.0/

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 491–496

Assessment: 37 yo M s/p total thryoidectomy for multi-nodular goiter that was impinging airway that appears to have a protected airway with no signs of compromise. ***Subjective***: No overnight events Able to void on own Pain well controlled with morphine Amoxicillin Unspecified Lamictal (Oral) (Lamotrigine) Rash; ... ***Objective***: icterus. Constricted pupils and PERRLA/EOMI. MMM. OP clear. Anterior neck C/D/I CARDIAC: Regular rhythm, normal rate. Normal S1, S2. No murmurs, rubs. ...
Summary: S/P Total Thyroidectomy

Table 1: An example of a clinical note, taken form 108627.txt from the training data.

Plan. The *Subjective* component are mainly from the patient's own description of the symptoms, the *Objective* component concludes structured clinical data such as blood test result, the *Assessment* component is the doctor's passive and active diagnoses, the *Plan* component lists the medical problems and the corresponding treatment plans. In this task, the training data includes the first three components , and a more detailed annotation for the *Plan* component is used as the problem summary, in which there are direct and indirect problems. Table 1 shows an example of the clinical note. We also notice that the size of task 1A's training data is similar to that in (Gao et al., 2022) (765 v.s. 768), but the size of the test set is larger (237 v.s. 92). To determine whether clinical note summarization is more extractive or abstractive. We conduct lowercase and compare the intersection of the n-gram tokens from both *Assessment* and *Summary* based on the training data, Figure 1 shows the number of clinical notes for different percentage of n-grams which appear in both the *Assessment* and *Summary*. We can see all of these three follow a long-tail like distribution, with uni-grams appearing more often in the *Summary* than bi-grams or tri-grams. We also compute the median value for the above n-gram overlap ratios, and get 0.22/0.20/0.13 for uni/bi/tri gram overlap between the *Assessment* and *Summary*, this also provides an estimated upper-bond for extractive models. In general, we anticipate that this task is more suitable with abstractive summarization modeling. Later experiments also validate our hypothesis.

4 Clinical Progress Note Summarization Modelling

We develop two types of clinical note summarization systems based on either rules or pre-trained language models. For rule based systems, it is largely relied on domain knowledge (e.g. UMLS) and expert information (e.g. doctors' pre-diagnosis). For pre-trained langague models, fine tuning them needs a large amount of high quality data. Even though nowdays pre-traied language models show superior performance on many NLP tasks, for this specific task we develop our own summarization rules without too much expert involvement.

4.1 Rule-based Summarization with UMLS

After some manual review of the *Summary* of the clinical notes, we have three findings: the *Summary* is very likely to include disease names and key findings and less likely to contain negated clinical terms, and key information appears more frequently in the first few sentences of the clinical note. Therefore, while developing the summarization rules, we consider three aspects for a clinical note: UMLS annotation, negation and sentence position. Given a clinical note and its *Summary*, we first apply a UMLS annotator MedCat (Kraljevic et al., 2021) to get the medical concepts and semantic types. Based on all the 756 summaries, we get the frequency of all the medical concepts and semantic types and filter out two lists: a key medical concept list C in which the frequency of the concepts is bigger than 50, a key semantic type list T of which the frequency of the semantic types is bigger than 90. Table 2 shows the list of the highly ranked semantic types. We also show the highly ranked concepts in the Appendix. Algorithm 1 shows the rule-based summarization process. We first use a SpaCy biomedical NER to recognize all the named entities, and filter out those entities of which there is no negation detected [3] and appear in either C or T. After the negation and UMLS filtering, we find there is around 0.9% clinical notes which output empty summaries. For these clinical notes, we simply take the very first sentence as the summary.

[3]We use negspaCy from https://spacy.io/universe/project/negspacy

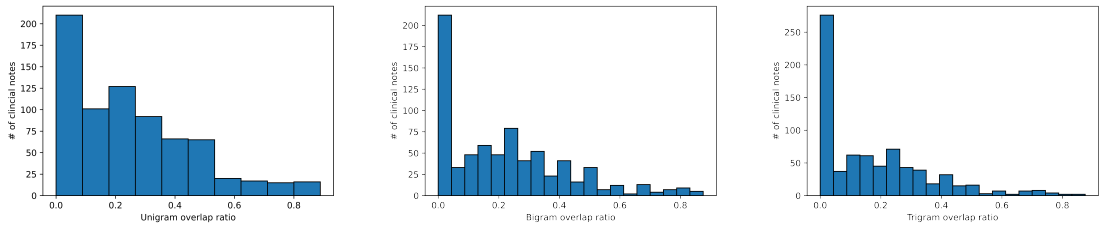

Figure 1: Each histogram shows the number of clinical notes for different ranges of overlapped n-gram percentage in the *Summary*. (a) uni-gram, (b) bi-gram, (c) tri-gram.

Semantic ID	Description
T047	Disease or Syndrome
T033	Finding
T046	Pathologic Function
T184	Sign or Sympton
T080	Qualitative Concept
T191	Neoplastic Process
T169	Functional Concept
T079	Teemporal Concept
T061	Therapeutic

Table 2: Highly ranked semantic types, all of these semantic types appear more than 90 times in the Summary column of the training data.

Algorithm 1 Rule-based clinical note summarization

Input: A clinical note D
Output: A Summary S

$S \leftarrow \emptyset$
$L \leftarrow \texttt{SapcyNER}(D)$
while $len(L) \neq 0$ **do**
  $s \leftarrow \texttt{pop}(L)$
  $c, t \leftarrow \texttt{MedCat}(s)$
  if s *is* `negated` **then**
    continue
  else
    if $c \in C$ `or` $t \in T$ **then**
      $S \leftarrow S + s$
    end
  end
end
if $len(S) < 1$ **then**
  `Add the first sentence of` D `to` S
end
return S

4.2 Fine-tuned Language Models

We fine tune three pre-trained language models on the provided training data, including BART (Lewis et al., 2019), T5 (Raffel et al., 2020) and GPT2 (Radford et al., 2019)[4]. BART and T5 are both denoising autoencoders for pretraining sequence-to-sequence models. BART is trained by corrupting text with an arbitrary noising function and learning a model to reconstruct the original text. T5 is also a transformer-based model that is pre-trained with token masking. We use the `bart-base` [5] and `T5-small` [6] due to their light implementation. Because of time limit, we did not conduct second time language model pre-training on MIMIC-III data but directly fine tune BART and T5 on the provided training data. Since GPT2 has an autoregressive decoder architecture, we use ClinicalBERT (Alsentzer et al., 2019)[7] as an encoder and fine tune both ClinicalBERT and GPT2 on the training data directly.

5 Experiments

Setup For both the rule-based and langague models, we only include the *Assessment* and *Summary* component, i.e., we drop the *Subjective* and *Objective* columns. The reason is that we do not find any significant performance difference without *Subjective* and *Objective*. We also follow (Gao et al., 2022) with some data augmentation by replacing those MedCat annotated text spans with their synonyms. Different from (Gao et al., 2022), for each clinical note, we randomly sample the MedCat annotated text spans and replace with their synonyms, in total we get another 38,250 augmented clinical notes. To save the training time,

[4] https://huggingface.co/gpt2
[5] https://huggingface.co/facebook/bart-base
[6] https://huggingface.co/t5-small
[7] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

we take the last checkpoint of the fine tuned T5 and continue training based on the augmented data set.

Result We use ROUGE-L (Lin, 2004), a widely used metric in summarization evaluation that based on n-gram overlap. Table 3 shows the RL-Precision, RL-Recall and RL-F1 results for all the submitted models on the blind test set. We also include the result of the best ranked system [8]. Our experiments validate previous results from (Gao et al., 2022) that T5 is the best language model for the task of clinical note summarization and achieves 0.208 Rouge-L F1 score. To our surprise, the rule-based system, BART and GPT2 fail on this task, the Rouge-L F1 scores of these three models are all below 0.15. The augmented T5 achieves 0.188 Rouge-L F1, we anticipate that the model is not fully trained on the augmented data set and may suffer from the catastrophic forgetting problem. We notice there is still a big gap between our fine tuned T5 and the best ranked system, of which the Rouge-L F1 is 0.327.

Human Analysis To further analyze why the rule based system fail on this task, we randomly sampled 50 clinical notes and asked a medical expert to read the output summaries from the rule-based and T5-based model and vote for the preferred ones. T5 got 60% of the vote and the rule system got the other 40%. Figure 2 shows an example of the output summary from the rule-based model and the fine tuned T5. It can be seen that in the clinical note there are many abbreviations and the true summary is quite abstractive. For a non-expert, it is really hard to distinguish the quality of the rule-based summary and the fine tuned T5 output. But in general we can see that the rule-based model extract some non-necessary tokens, such as *obesity* and *oxygen*, which are not the key information in the true summary. In contrast, all the generated terms from the fine tuned T5, such as *SOB* (short of breath) and *CAD* (coronary artery disease) are important symptom or disease related terms.

6 Conclusion

In this paper, we summarized the models we developed for the BioNLP2023 workshop task 1A. In the age of large language models, we developed a simple rule based system for clinical note summarization, which is extractive and based on UMLS. We also fine tuned three pre-trained language models to the clincal summarization task. Experiment results based on Rouge-L show that the fine tuned T5 model achieves the best performance. Further human analysis also validated the superiority of the fine tuned T5 over the rule based system.

Model	RL-P	RL-R	RL-F1
Rule	0.082	0.034	0.043
BART	0.064	0.031	0.039
ClincalBERT+GPT2	0.121	0.180	0.133
T5	0.286	0.191	0.208
T5-augmented	0.284	0.158	0.188
Best ranked system	**0.416**	**0.305**	**0.327**

Table 3: Model comparison for clinical note summarization.

Assessment: 66 yo F with PMH of morbid obesity, PAF, DM2, COPD on 2L home oxygen, CAD and recent diagnosis of PE who presents from OSH with SOB and fevers.
True summary: SOB: Most likely explnanation currently is aspiration; GNR Bacteremia; COPD; PE; afib; CAD; DM2
Rule-based summary: obesity; DM2; COPD; oxygen; CAD; PE; SOB
T5 summary: SOB; CAD; DM2; CAD; DM2; COPD

Figure 2: Example summaries from rule-based model and fine tuned T5.

Limitations

There are a few limitations in our work: First, the rules developed are totally based on frequency filtering and not further checked by medical experts, we are not sure whether there are any hidden template or patterns in the clincal note summary. Second, due to time limitation, we did not conduct second time pre-training for the language models, T5 was originally trained from generic text of which the genre is quite different from the clinical domain. 765 training examples may not be enough for the model to learn. Third, we would like to give a full test of the more recent large language model (e.g. ChatGPT[9]), but we cannot fine tune it with the open free API.

Ethics Statement

All the experiment data is from PhysioNet [10]. To get full access to the training and test data in this

[8] The result of the best ranking system is taken from https://codalab.lisn.upsaclay.fr/competitions/12388 on 28 April 2023.

[9] https://chat.openai.com/

[10] https://physionet.org/content/bionlp-workshop-2023-task-1a/1.0.0/

task, it is required to get the training for CITI Data or Specimens Only Research [11].

References

Griffin Adams, Emily Alsentzer, Mert Ketenci, Jason Zucker, and Noémie Elhadad. 2021. What's in a summary? laying the groundwork for advances in hospital-course summarization. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2021, page 4794. NIH Public Access.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, WA Redmond, and Matthew BA McDermott. 2019. Publicly available clinical bert embeddings. *NAACL HLT 2019*, page 72.

Yu-Neng Chuang, Ruixiang Tang, Xiaoqian Jiang, and Xia Hu. 2023. Spec: A soft prompt-based calibration on mitigating performance variability in clinical notes summarization. *arXiv preprint arXiv:2303.13035*.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, Dongfang Xu, Matthew M. M. Churpek, and Majid Afshar. 2022. Summarizing patients' problems from hospital progress notes using pre-trained sequence-to-sequence models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2979–2991, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yanjun Gao, Dmitry Dligach, Timothy Miller, Matthew M. Churpek, and Majid Afshar. 2023. Overview of the problem list summarization (probsum) 2023 shared task on summarizing patients' active diagnoses and problems from electronic health record progress notes. In *Proceedings of the 22nd Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Neel Kanwal and Giuseppe Rizzo. 2022. Attention-based clinical note summarization. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, pages 813–820.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, et al. 2021. Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. *Artificial intelligence in medicine*, 117:102083.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jennifer Liang, Ching-Huei Tsou, and Ananya Poddar. 2019. A novel system for extractive clinical note summarization using ehr data. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 46–54.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Wen-wai Yim and Meliha Yetisgen-Yildiz. 2021. Towards automating medical scribing: Clinic visit dialogue2note sentence alignment and snippet summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 10–20.

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. *arXiv preprint arXiv:2109.12174*.

[11] https://physionet.org/content/bionlp-workshop-2023-task-1a/view-required-training/1.0.0/#1

A Appendix

The following list shows the key UMLS concepts which are used for our rule-based summarization model.

- C0020649 : Hypotension. Definition: Abnormally low BLOOD PRESSURE that can result in inadequate blood flow to the brain and other vital organs. Common symptom is DIZZINESS but greater. Semantic Types: Findings
- C0004238: Atrial Fibrillation. Definition: Abnormal cardiac rhythm that is characterized by rapid, uncoordinated firing of electrical impulses in the upper chambers of the heart (HEART ATRIA) Semantic Types: Disease or Syndrome
- C0020538: Hypertensive disease. Definition: Persistently high systemic arterial BLOOD PRESSURE. Based on multiple readings (BLOOD PRESSURE DETERMINATION), hypertension is currently defined as. Semantic Types: Disease or Syndrome.
- C0022660: Kidney Failure, Acute (C0022660) ** Definition: Sudden and sustained deterioration of the kidney function characterized by decreased glomerular filtration rate, increased serum creatinine. Semantic Types: Disease or Syndrome
- C0242184: Hypoxia. Definition: Sub-optimal OXYGEN levels in the ambient air of living organisms. Semantic Types: Pathologic Function
- C1145670: Respiratory Failure. Definition: A severe form of respiratory insufficiency characterized by inadequate gas exchange such that the levels of oxygen or carbon dioxide cannot be ... Semantic Types: Disease or Syndrome
- C1956346: Coronary Artery Disease. Definition: Pathological processes of CORONARY ARTERIES that may derive from a congenital abnormality, atherosclerotic, or non-atherosclerotic cause. Semantic Types: Disease or Syndrome
- C0015967: Fever. Definition: An abnormal elevation of body temperature, usually as a result of a pathologic process. Semantic Types: Sign or Symptom
- C0332148: Probable diagnosis. Semantic Types: Finding
- C0023518: Leukocytosis. Definition: A transient increase in the number of leukocytes in a body fluid. Semantic Types: Disease or Syndrome
- C0036690: Septicemia. Definition: Systemic disease associated with the presence of pathogenic microorganisms or their toxins in the blood. Semantic Types: Disease or Syndrome
- C0278061: Abnormal mental state. Definition: A reduction in the subjective feeling of mental well being. Semantic Types: Mental or Behavioral Dysfunction
- C0011849: Diabetes Mellitus. Definition: A heterogeneous group of disorders characterized by HYPERGLYCEMIA and GLUCOSE INTOLERANCE. Semantic Types: Disease or Syndrome
- C0002871: Anemia (C0002871)