LAILab at Chemotimelines 2024: Finetuning sequence-to-sequence language models for temporal relation extraction towards cancer patient undergoing chemotherapy treatment

Shohreh Haddadan[1], Tuan-Dung Le[1,2], Thanh Duong[1,2], Thanh Q. Thieu[1,2],
[1]Moffitt Cancer Center and Research Institute, USA
[2]University of South Florida, USA
{shohreh.haddadan, tuandung.le , thanh.duong, thanh.thieu}@moffitt.org

Abstract

In this paper, we report our effort to tackle the challenge of extracting chemotimelines from EHR notes across a dataset of three cancer types. We focus on the two subtasks: 1) detection and classification of temporal relations given the annotated chemotherapy events and time expressions and 2) directly extracting patient chemotherapy timelines from EHR notes. We address both subtasks using Large Language Models. Our best-performing methods in both subtasks use Flan-T5, an instruction-tuned language model. Our proposed system achieves the highest average score in both subtasks. Our results underscore the effectiveness of finetuning general-domain large language models in domain-specific and unseen tasks.

1 Introduction

Patient health records contain a wealth of information that can offer valuable insights to healthcare professionals and researchers, aiding in the enhancement of diagnosis, treatment, and disease prevention. Cancer patients often undergo lengthy treatment regimens, resulting in extensive electronic health record (EHR) documentation over time. The sheer volume of data available to healthcare providers is substantial, making manual curation impractical and cost-prohibitive.

A crucial aspect of cancer patient records is their chemotherapy treatment status documentation. Automatically extracting information regarding the timelines of chemotherapy events offers several advantages, including the ability to evaluate treatment efficacy across various cancer types. This automated extraction process also facilitates the creation of concise summaries for future attending physicians.

Two main tasks have been defined and addressed in association with temporal relation extraction from clinical notes: DocTimeRel and TLINK detection and classification. The first task is to identify and classify the relation between events in an EHR note and the creation time of the document. TLINK detection and classification identify relations between event mentions and time expressions in EHR notes.

In this paper, we deal with the latter, the temporal relation extraction on a dataset of three cancer types. The shared task defines two subtasks. Subtask one aims at discerning a temporal relationship between a pair, consisting of a chemotherapy event and a time expression, subsequently classifying this relationship into one of the following categories: CONTAINS, BEGINS-ON, or ENDS-ON. In the second subtask, the only given input is the patient notes. The desired output for both subtasks is patient-level chemotherapy timelines. For detailed information on the definition of the subtasks, baseline methods, dataset, and evaluation criteria, see (Yao et al., 2024).

We approach both subtasks using large language models (LLMs). For the first subtask, we reformulate the relation classification problem into a text generation task and benefit from instruction-tuned language models to predict the relation. In the second subtask, we experiment with a sequence-to-sequence fine-tuning method with relations transformed into target sequences using a triplet linearization algorithm and also a pipeline method consisting of a rule-based event and time expression module and our best-performing model on the first subtask to identify and classify the pairwise relations.

We achieved the highest average scores on the test results as announced by the organizers (Yao et al., 2024).

In the following, we describe how we have utilized LLMs in detection, classification, and the end-to-end approach to chemotherapy timeline extraction from clinical notes.

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 382–393

Instruction:
An EVENT is anything that is relevant on the clinical timeline.
Temporal expressions (TIME) are discrete references to time.
Temporal relations link an EVENT and TIME.
The set of temporal relations is CONTAINS, ENDS-ON, BEGINS-ON, NO-RELATION.
Given an input text describing the relationship between an EVENT and TIME, extracts the relationship between them.
The markers <t> and </t> delineate the TIME entity.
The markers <e> and </e> delineate the EVENT entity.
Note: Your output must only be the relation of the two given entities and must follow the format: "Relation: <One of the above listed relations>"

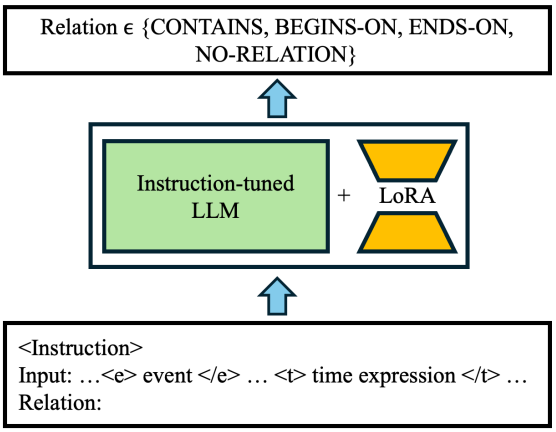

Figure 1: Low-rank adaptation instruction fine-tuning for Subtask 1

2 Methodology

2.1 Subtask1

With the chemotherapy events and time expressions in each patient's note already provided by the organizer, the first subtask aims to identify temporal relations between them and subsequently generate patient-level timelines.

Prior to training a model, we need to prepare the dataset to train the model for the temporal relation classification task. The annotated relations with their respective pair of events and time expressions in the gold standard training/development dataset are added as positive instances. We create negative instances labeled as NO-RELATION by pairing events and time expressions within a patient note with no temporal relations in the gold-standard dataset.

However, incorporating every potential negative instance would lead to a significant imbalance in the training dataset as well as additional computational costs for training and inferencing the model. To mitigate this, we exclude instances where the positional distance between the event mentions and time expressions in the EHR note exceeds a maximum number of characters. Table 1 reports the maximum distance and number of NO-RELATION label instances added to the dataset. With this empirical observation, we set the maximum distance to 250 characters. We also create a heuristic rule that any pairs with a distance greater than the threshold are automatically predicted as NO-RELATION during inference on the test set. Applying this rule to the test set reduces the number of possible pairs from 12762 to 3042, thus enabling a more computationally efficient inference process.

During preprocessing, we first employ the "mimic" model from the Stanza library, developed by (Zhang et al., 2021), for sentence segmentation. Then, we construct the context for the input sequences using two different approaches: concatenated context and bounded context. If the event and the time expression in the pair occur within the same sentence, both methods consider the sentence as the context. Otherwise, if the event and time expression are located in different sentences, the two sentences are concatenated to form the concatenated context. In the bounded context method, any sentence occurring between these two sentences is also included in the context. In addition, we add markers denoted by <e> followed by </e>, and <t> followed by </t> to respectively delineate events and time expressions.

We reformulate the temporal relation classification task as a generation task by finetuning a large language model to directly generate a label from the predefined set of relation types: CONTAINS, BEGINS-ON, ENDS-ON, NO-RELATION. We prepend the instruction describing the task to each input context. This method conditions the model to generate the relation type immediately following an anchor prompt "Relation:". Figure 1 illustrates our approach to tackling the first subtask, including our model's input and expected output. In the instruction, we use the definition of events, time expressions, and temporal relations provided in the data descriptions of the shared task. Our instruction format leverages the prompt structure used in relation extraction tasks, as described by (Lai et al., 2023). We also experiment with finetuning the model without adding the task instruction to the input contexts.

During our preliminary experiments, we fine-

Split	Cancer type	Gold relation pairs#	Max character distance	No relation pairs #
Train	brca	455	99	381
	mela	48	218	35
	ovca	494	143	336
Dev	brca	113	213	132
	mela	201	144	191
	ovca	226	173	220

Table 1: Maximum (character) distance between event and time mentions of relation pairs in the gold standard dataset used as a threshold to reduce the number of NO-RELATION pairs. The number of gold relations is provided for comparison.

tune three instruction-tuned models: Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), Flan-T5-xxl 11B (Chung et al., 2022) and Llama-2-13B-chat (Touvron et al., 2023).

Flan-T5-xxl consistently achieved superior performance on the development set compared to the other two models. Thus, we use Flan-T5-xxl for further experimentation in this subtask.

2.2 Subtask2

As Yao et al. (2024) describe, in the second subtask, the input to the system is only the patient's EHR notes. Therefore, an end-to-end system that integrates identifying chemotherapy events and time expressions to extract the final chemotherapy patient-level timeline is required. We consider two different approaches to address this subtask.

In the first approach, we train a sequence-to-sequence model with input snippets from the EHR notes. The output is sequences containing the temporal relations (each a triplet of <event, relation type, time expression>) found in that snippet. The training objective is to simultaneously identify the events and time expressions in the context of each sentence in the EHR note and to detect and classify the relation as CONTAINS, BEGINS-ON, or ENDS-ON. We use the annotated data for the first subtask to train the models and evaluate our models on the development set provided for the second subtask.

We consider the context of each sentence to be its neighboring sentences (one preceding and one succeeding) joined by the separator token of the corresponding tokenizer as defined in equation (1).

$$Context(s_i) = s_{i-1} + [SEP] + s_i + [SEP] + s_{i+1} \quad (1)$$

Huguet Cabot and Navigli (2021) have neatly introduced a triplet linearization algorithm for generating target sequences incorporating one or more relations between entities. We adopt this algorithm to transform the annotated temporal relations to target sequences.

Our approach differs from Huguet Cabot and Navigli (2021)'s approach in several ways. Firstly, their approach is to identify more than 200 relation types; thus, contrary to our setting, they are not limited to a restricted set. We add the relation types (CONTAINS, BEGINS-ON, ENDS-ON) to the special tokens of the tokenizer so they are not split during the tokenization process and the model learns them as defined in the target sequences. Since the events in our problem settings are domain-specific, we observed that the approach used in Huguet Cabot and Navigli (2021) identifies any event (not only the chemotherapy events) as a chemotherapy event after training. To prevent the generation of false positive events, we include additional chemotherapy events annotated in the gold standard data set, which are not in any relation with a time expression, to the training set. Similarly, to create negative instances, we add the input sequences that have no annotation of chemotherapy events, time expressions, or relations to the training data. Figure 2 shows different input sequence and their corresponding target sequence.[1]

We then experiment with finetuning various versions of two pre-trained models with the encoder-decoder structure, which have proven to perform well for sequence-to-sequence tasks, namely BART and Flan-T5. The reasoning behind choosing BART is that it is trained for sequence-to-sequence tasks and has proven to perform well on sequence-generation tasks. We chose Flan-T5 to test the effectiveness of this instruction-tuned model on an unseen task. In this subtask, we do not add instructions to input sequences while finetuning Flan-T5. We conduct experiments on various available model sizes for BART and Flan-T5.

In the second approach, we use a pipeline method that consists of two steps: the first step extracts the chemotherapy events and time expres-

[1]To abide by the terms of the data agreement, we refrain from quoting exact snippets from the EHR notes. The examples are altered and, therefore, might not be medically accurate.

Input Sequence$_1$	They underwent surgery. On day of admission, they had their first dose of Taxol. Their blood glucose was 456.
Target Sequence$_1$	<triplet> day of admission <subj> Taxol <obj> CONTAINS
Input Sequence$_2$	Vital signs are stable. Culture results significant cancer, currently getting chemotherapy. They present for evaluation today.
Target Sequence$_2$	<triplet> currently <subj> chemotherapy <obj> CONTAINS
Input Sequence$_3$	Patient seen on 04/12. They received first dose of aflibercept today and second dose 05/16 prior to admission for high dose. Transferred to unit.
Target Sequence$_3$	<triplet> aflibercept <subj> today <obj> BEGINS-ON <triplet> 05/16 <subj> aflibercept <obj> CONTAINS
Input Sequence$_4$	No known medicinal allergies. They were initiated on the ipilimumab arm. They continue on the recommended regimen.
Target Sequence$_4$	<triplet> <subj> ipilimumab <obj>
Input Sequence$_5$	Malignant melanoma of other specified site. Patient here for cycle #2 of IL-2, on study with aflibercept. follow electrolytes and renal function
Target Sequence$_5$	<triplet> <subj> IL-2 <obj> <triplet> <subj> aflibercept <obj>
Input Sequence$_6$	Patient was seen and examined. History and physical exam were reviewed. I agree with physical findings.
Target Sequence$_6$	<triplet> <subj> <obj>

Figure 2: The input sequences are the contexts, including a sentence and its preceding and succeeding sentence in the EHR note joined by the separator token of the corresponding tokenizer. Target sequences are the linearized triplets taken from the gold standard annotations. Following the encoding in Huguet Cabot and Navigli (2021), <triplet> marks the start of a new temporal relation with a new head entity, followed by the tokens representing the head entity in the input text; <subj> marks the end of the head entity and the start of the tail entity's tokens; <obj> marks the end of the tail entity and the start of the relation type between the head and tail entity. The head/tail entities can be either a chemotherapy event or a time expression depending on their relative position in the text.

sions, and in the second step, we utilize our best-performing model on the first subtask to detect and classify the relations between pair of events and time expressions. We extract the time expressions using the Python wrapper for Stanford CoreNLP's SUTime Java library developed by Manning et al. (2014)[2]. We utilize a rule-based system with a predefined dictionary for the event extraction task. We compile a list of chemotherapy events from three different sources: 1) the baseline system[3]. 2) all chemotherapy events extracted from the training set, and 3) all the cancer drugs mentioned on the Cancer Research UK website[4].

2.3 Finetuning process

Our approach uses the Huggingface[5] implementation of the Seq2SeqTrainer to finetune trained models.

In the first subtask, we set the maximum length of the input as 450 tokens and the maximum target length as 10 tokens to fit the instruction. We finetune Flan-T5-xxl model using LoRA (Hu et al., 2021) for 10 epochs, employing early stopping with a patience of 3 epochs.

In the second subtask, we set the maximum length of the input as 256 tokens and the maximum target length as 32 tokens. We then pad input and target sequences to maximum length with the pad token of the tokenizer specific to the model. We run the finetuning for 10 epochs in the BART setting and 5 epochs in Flan-T5 setting. The parameter efficient module (LoRA) was enabled while finetuning Flan-T5 models for this subtask. For more details on the models, see Appendix A.1.

For both finetuning experiments, we used the implementation of LoRA in Huggingface library. Parameters for LoRA are set to α = 32, dropout = 0.05, and r = 16 and are added to $[q, k, v, o]$ layers in both tasks. Appendix A.2 briefly describes LoRA.

2.4 Preparing data for evaluation

Most of the time expressions in the EHR notes are relative and must be normalized using the document time (DOCTIME). The document time in the first subtask can be extracted from the gold standard annotated data or the headers of each patient EHR note. In the case of the second subtask, only the latter is feasible due to the absence of gold standard annotations. The headers of the patient records are provided in a standard format, so the document time can be precisely extracted using regular expressions.

To normalize relative time expressions such as "two weeks ago", "today", "currently", we use the timenorm library (Xu et al., 2019). We discard the extracted relations for which timenorm fails to

[2] https://pypi.org/project/sutime

[3] https://github.com/HealthNLPorg/chemoTimelinesBaselineSystem/tree/main/timelines/instance-generator/src/user/resources/org/apache/ctakes/dictionary/lookup/fast/bsv

[4] https://www.cancerresearchuk.org/about-cancer/treatment/drugs

[5] https://huggingface.co/

Cont.	Inst.	brca			mela			ovca		
		F1	RF1	TF1	F1	RF1	TF1	F1	RF1	TF1
Bound	No	0.893	**0.992**	0.941	0.922	0.938	**0.887**	0.879	0.968	0.852
Bound	Yes	**0.922**	0.980	**0.962**	**0.960**	**0.977**	**0.887**	**0.916**	**0.987**	0.793
Concat	No	0.913	0.980	0.937	0.898	0.916	**0.887**	0.890	0.968	**0.871**
Concat	Yes	0.919	0.967	0.918	0.934	0.954	**0.887**	0.893	0.960	0.810

Table 2: Results for the first subtask on the development set. The terms F1 and RF1 represent the F1-score and relaxed F1-score of our classification model, respectively. TF1 is the official F1-score for the final timelines calculated using the evaluation system.

normalize the time expression. Examples of such time expressions include "at this time", "January 10 or 11", "05/2012" , "day one" , "16-09", etc.

For both subtasks, we use the baseline system provided by Yao et al. (2024) for de-duplication and creation of final timelines.[6]

As an extra step in the pipeline approach to the second subtask, we manually omitted some events and time expressions from the results of the rule-based systems. Examples of such omissions are "continues" event (which appears in the train set) and time expressions "1842", "1255" and "1000".

3 Results

We use the evaluation system provided by (Yao et al., 2024) for both subtasks on the development set.[7] The evaluation script receives the gold standard timelines, and the system prediction for all patients in each cancer type as input.

All the experiments on the development set have been executed before the test set results were announced.

3.1 Subtask1

In addition to reporting the timeline score on the development set using the organizer's evaluation system, we also evaluate our model's performance on the pairwise temporal classification task (Table 2). We implement two metrics: micro F1 and relaxed micro F1. CONTAINS and BEGINS-ON, CONTAINS and ENDS-ON are interchangeable in the relaxed F1-score computation.

Finetuned Flan-T5-xxl with instruction and bounded context achieved the highest scores on almost all metrics. Finetuning bounded context shows a marginal improvement in relaxed micro F1 compared to the concatenated context. This suggests that incorporating sentences between sentences containing event and time expression might be beneficial for classifying NON-RELATION pairs.

Our classification model scores do not correlate well with timeline scores. For instance, in ovarian cancer results, fine-tuned Flan-T5-xxl bounded context and instruction achieves the highest F1-score on the classification task but the lowest timeline score. We suspect that this difference originates from official results being based on average macro F1 score across all patients. Further reasons might be related to the errors of the post-processing steps in creating the final patient timelines, such as the normalization of time expressions and the de-duplication process. We select three submissions with the highest average F1-score of F1-scores, relaxed F-scores and final timeline F-scores for all cancer types as presented in Table 2.

Our submission outperformed the baseline for breast cancer, ovarian cancer, and the average score. It achieved the same score as the baseline system for melanoma cancer (Table 3).

3.2 Subtask2

The end2end approach with Flan-T5-xxl + LoRA achieves the highest results across all other methods and the baseline system results for melanoma and ovarian cancer as shown in Table 4. For breast cancer, on the other hand even though this method performs best among other implemented methods, it does not surpass the baseline system results on the development set.

Considering the relaxed setting, Flan-T5-xxl + LoRA has achieved the highest precision rate across all cancer types. However, the methods that extract event types using a rule-based or dictionary-based system (baseline system and the pipeline approach) have gained higher recall scores in the

[6] https://github.com/HealthNLPorg/chemoTimelinesBaselineSystem

[7] https://github.com/HealthNLPorg/chemoTimelinesEval

	Method	brca	mela	ovca	Average score
Subtask 1	Baseline system	0.93	**0.87**	0.88	0.89
	Flan-T5-xxl + bound context + instruction	**0.96**	**0.87**	0.88	**0.90**
	Flan-T5-xxl + bound context	0.95	0.85	**0.89**	**0.90**
	Flan-T5-xxl + concat context	0.95	0.84	**0.89**	**0.90**
	Highest score on the leader board	0.96	0.87	0.89	0.90
Subtask 2	Baseline system	0.59	0.43	0.71	0.58
	End2end BART-large	0.52	0.57	0.59	0.56
	End2end Flan-T5-xxl + LoRA	0.62	**0.74**	**0.74**	**0.70**
	Pipeline system	0.53	0.38	0.49	0.47
	Highest score on the leader board	0.68	0.74	0.74	0.70

Table 3: Evaluation published by the organizers for our submission on the held-out test set

same setting. The low score in the strict evaluation setting for Flan-T5 is due to its failure to identify ENDS-ON relations in any cancer type, possibly because of the label's low frequency in the training set. See Appendix B for detailed results on precision and recall.

We chose to submit the results of the end2end method with both BART and Flan-T5 with the highest scores, which are the largest models we experimented with, namely BART-large and Flan-T5-xxl and the pipeline approach as our third submission.

We first performed a sentence tokenization step on the test data and extracted the contexts as input sequences. We used the models to infer target sequences. The results of these three approaches on the test data provided by the organizers are presented in Table 3. The end2end approach using the pre-trained Flan-T5-xxl model with LoRA exceeds in all evaluations except for the breast results. Albeit, the breast cohort results surpass the test set's baseline score contrary to the experiments on the development set. The average score on this approach gains the highest score among other submissions, as reported by the organizers.

Method	brca	mela	ovca
Baseline system	**0.857**	0.456	0.329
Pipeline Approach	0.529	0.511	0.470
End2End BART-large	0.700	0.618	0.496
End2End Flan-T5-xxl	0.749	**0.720**	**0.647**

Table 4: Evaluation for the second subtask on the development set.

4 Error Analysis

Since the gold standard timeline and annotations on the test set have not been released to enable future editions of the task, we will provide the error analysis on the results of the development set.

4.1 Subtask1

Our best model, Flan-T5-xxl finetune bounded context with instruction, achieved a low error rate of 20 incorrect predictions out of 1,083 tested pairs. Possible error sources include misspellings, potentially missing or spurious annotations, or unclear or complex contexts. Complex contexts occur when notes include tables that have lost their structures in the plain text files. We list some examples of mispredictions below.

- Misspelling: "Condition **<t>yesterdat</t>** appeared improved with treatment, and **<e>chemo</e>** cycle discontinued.", Label: ENDS-ON, Predict: NO-RELATION.
- Missing annotation: "Patient with metastatic melanoma enrolled in protocol and received first dose of **<e>aflibercept</e>** 9/4 and second dose **<t>09/18</t>** prior to admission for high dose IL2 (first cycle)Thus far has received 9/12 planned doses.", Label: NO-RELATION, Predict: CONTAINS.
- Spurious annotation: "Patient enrolled in protocol and received first dose of alibercept 9/4 and second dose **<t>09/18</t>** prior to admission for high dose **<e>IL2</e>** (first cycle)", Label: CONTAINS, Predict: NO-RELATION.
- Unclear context: "Cycle #2 was initiated on **<t>September 10 , 2011</t>**; however, the patient had a severe reaction during the **<e>paclitaxel</e>** infusion.", Label: CONTAINS, Predict: NO-RELATION.

4.2 Subtask2

One source of error in subtask2 is the emergence of medical events or drugs as output events that

are not particularly chemotherapy events such as "radiation", "iv decadron", "bolus", "anti-vegf antibody", "augmentin" and so on. The following example shows one of the incident where our best performing model incorrectly identifies "radiation" as a chemotherapy event in temporal relation CONTAINS with "June 1st". [8]

- ... she is undergoing consultation with Dr. X for possible **radiation** therapy on June 1st.

We noticed that balancing the negative instances with the positive examples of temporal relations worsens this problem. Thus, we keep all negative instances in the training set to improve the identification of chemotherapy events. These negative instances include the ones containing events/time expressions but no relations, for example, target sequences 4 and 5 in Figure 2. And also the ones with no events and no time expressions, for example, input sequence 6 in Figure 2. We suspect this problem is caused by the bias of the pre-trained model in identifying all entities beyond the chemotherapy events. This approach improved the results; however, it's not completely resolved. It can further be alleviated either manually or by applying a post-processing filter created by experts to only keep the temporal relations with chemotherapy drugs/treatments.

In examining the distribution of relation types across various cancer types within the development set for the second subtask, we observed an imbalance in the dataset. Specifically, the ENDS-ON relation type was found to occur with frequencies of approximately 30%, 2%, and 14% concerning all chemotherapies within the final gold timelines for breast cancer, melanoma, and ovarian cancer, respectively. Given our approach's reduced accuracy in identifying the ENDS-ON relation type, this discrepancy explains the lower accuracy observed compared to the baseline system specifically concerning breast cancer within both the development and potentially the test set (Assuming the distribution of relation types on the test set is close to the distribution on the development set).

Another source of the model's confusion is the chemotherapy events that were not annotated in the training dataset. The first example was identified as a "chemotherapy" event in CONTAINS temporal relation with time expression "2003" and the second as "docetaxel", BEGINS-ON, "oct 3rd" by our end2end model, however, we do not find the equivalent of this chemotherapy event instance in the annotated development set. In order to resolve this particular error, we would need further information about the annotation rules.

- History of Present Illness: Patient was diagnosed with disease in 2003 and treated with surgery, chemotherapy, and radiation per the patient.

- Patient says they are now taking docetaxel with 1st dose Oct 3rd and second due in early november.

We can also associate a fraction of errors to the normalization errors originating from the *timnorm* library, for example, in cases where time expressions containing two-digit years are inaccurately resolved to the 1900s.

5 Related work

Numerous studies focus on annotation (O'Gorman et al., 2016; Wang et al., 2022; Alsayyahi and Batista-Navarro, 2023), detection and classification (Lim et al., 2023; Huang et al., 2023) of temporal relations in the general domain.

In the medical domain, temporal relation extraction also received attention for its benefits in longitudinal studies of medication, treatments, and diseases, as well as in summarizing clinical notes for physicians' further reference. THYME annotation guidelines and corpus (Styler IV et al., 2014) and its extension (Wright-Bettner et al., 2020) is a considerable effort in the specification of process of temporal relation annotation process in clinical narratives based on ISO-TimeML (Pustejovsky et al., 2010).

Prior to the introduction of transformer-based language models a few studies approached various tasks of temporal relation extraction problem with feature-based supervised machine learning algorithms and sequential neural networks (Xu et al., 2013; Lee et al., 2016; Alfattni et al., 2020, 2021). Moreover, Lin et al. (2018) utilized unlabeled data by self-training neural networks in clinical temporal relation extraction.

After the rise of transformer-based models, temporal relation extraction from clinical notes also benefited from this significant development in NLP methods using models such as BERT (Lin et al., 2020; Zhou et al., 2021), BioBERT and BART (Wright-Bettner et al., 2020) for clinical text representation. Lin et al. (2021) continue training BERT using a masking method called entity-centric masking strategy, where they use the MIMIC III dataset

[8] Examples in this section have been altered to abide by the data use agreement.

as their training data. Their results on temporal relation extraction shows improvements on baselines using the model pretrained using this approach.

Most end-to-end systems for temporal relation extraction in the clinical domain have been tackled using a pipeline approach consisting of modules for event and time expression extraction and pairwise temporal relation detection and classification. Dligach et al. (2022) on the other hand, explore the use of sequence-to-sequence models in extracting temporal relations from text. They experiment with various input/output representations and adopt those representations, which enable the reconstruction of the snippets with several relations and repetitive event names in a text snippet. They report this approach's results utilizing different sequence-to-sequence LLMs such as BART and T5. Miller et al. (2023) approach temporal relation extraction problem as an end-to-end task without given events and time expressions using a combination of domain-specific pre-trained language model PubmedBERT (Gu et al., 2021) and a multi-headed attention classifier on THYME2 dataset (Wright-Bettner et al., 2020).

Bethard et al. (2016, 2017) organized previous shared tasks to incentivize the research on temporal relation extraction from clinical notes.

6 Conclusions

This paper presents our effort in participating in the Chemotimeline shared task. We apply an instruction finetuning method for temporal relation detection and classification and a sequence-to-sequence approach for extracting timelines directly from EHR notes to solve the first and second subtasks. Our approach, leveraging the power of general-domain Large Language Models and further finetuning them with parameter-efficient methods, secured the highest average scores across the different cancer types for both subtasks. The results of our approach using Flan-T5-xxl + LoRA underscore the potential of instruction finetuning in enhancing the capabilities of LLMs for unseen natural language understanding and generation tasks, even on domain-specific data. In future work, we aim at augmenting the data for low-frequency relation types and also harnessing the power of provided unlabeled data to continue pre-training Large Language models and to investigate the effect on the results of extracting temporal relations from cancer patient EHR notes.

Limitations

There are several limitations to our experiments.

Firstly, our experiments were bounded by computational resource limitations. Specifically, our experiments employed the Flan-T5 model with parameter-efficient techniques due to constraints in available computational power on shared hardware and time. This limitation prevents us from comparing our methodology with implementations of Flan-T5 without LoRA approach. Secondly, we do not test our experiments on other datasets since annotated data in the medical domain on such a specific task is extremely scarce. Thus, we cannot claim that our results will be as high on different datasets. Moreover, since our method is fine-tuned on the provided data, practical use and release of the model are legally bound by the data agreement usage. Finally, our method uses a deep learning approach and, therefore, is limited by the explainability and interpretability constraints of such techniques.

References

Ghada Alfattni, Niels Peek, and Goran Nenadic. 2020. Extraction of temporal relations from clinical free text: A systematic review of current approaches. *Journal of Biomedical Informatics*, 108:103488.

Ghada Alfattni, Niels Peek, and Goran Nenadic. 2021. Attention-based bidirectional long short-term memory networks for extracting temporal relationships from clinical discharge summaries. *Journal of Biomedical Informatics*, 123:103915.

Sarah Alsayyahi and Riza Batista-Navarro. 2023. TIMELINE: Exhaustive annotation of temporal relations supporting the automatic ordering of events in news articles. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16336–16348, Singapore. Association for Computational Linguistics.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. SemEval-2016 task 12: Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1052–1062, San Diego, California. Association for Computational Linguistics.

Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. SemEval-2017 task 12: Clinical TempEval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572, Vancouver, Canada. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

Dmitriy Dligach, Steven Bethard, Timothy Miller, and Guergana Savova. 2022. Exploring text representations for generative temporal relation extraction. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 109–113, Seattle, WA. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Quzhe Huang, Yutong Hu, Shengqi Zhu, Yansong Feng, Chang Liu, and Dongyan Zhao. 2023. More than classification: A unified framework for event temporal relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9631–9646, Toronto, Canada. Association for Computational Linguistics.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Hee-Jin Lee, Hua Xu, Jingqi Wang, Yaoyun Zhang, Sungrim Moon, Jun Xu, and Yonghui Wu. 2016. UTHealth at SemEval-2016 task 12: an end-to-end system for temporal information extraction from clinical notes. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1292–1297, San Diego, California. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Stanley Lim, Da Yin, and Nanyun Peng. 2023. LEAF: Linguistically enhanced event temporal relation framework. In *Proceedings of the 2nd Workshop on Pattern-based Approaches to NLP in the Age of Deep Learning*, pages 6–19, Singapore. Association for Computational Linguistics.

Chen Lin, Timothy Miller, Dmitriy Dligach, Hadi Amiri, Steven Bethard, and Guergana Savova. 2018. Self-training improves recurrent neural networks performance for temporal relation extraction. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 165–176, Brussels, Belgium. Association for Computational Linguistics.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2021. EntityBERT: Entity-centric masking strategy for model pretraining for the clinical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 191–201, Online. Association for Computational Linguistics.

Chen Lin, Timothy Miller, Dmitriy Dligach, Farig Sadeque, Steven Bethard, and Guergana Savova. 2020. A BERT-based one-pass multi-task model for clinical temporal relation extraction. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 70–75, Online. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Timothy Miller, Steven Bethard, Dmitriy Dligach, and Guergana Savova. 2023. End-to-end clinical temporal information extraction with multi-head attention. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2023, page 313. NIH Public Access.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging

annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas. Association for Computational Linguistics.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An international standard for semantic annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

William F. Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 926–941, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kristin Wright-Bettner, Chen Lin, Timothy Miller, Steven Bethard, Dmitriy Dligach, Martha Palmer, James H. Martin, and Guergana Savova. 2020. Defining and learning refined temporal relations in the clinical narrative. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 104–114, Online. Association for Computational Linguistics.

Dongfang Xu, Egoitz Laparra, and Steven Bethard. 2019. Pre-trained contextualized character embeddings lead to major improvements in time normalization: a detailed analysis. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 68–74, Minneapolis, Minnesota. Association for Computational Linguistics.

Yan Xu, Yining Wang, Tianren Liu, Junichi Tsujii, and Eric I-Chao Chang. 2013. An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):849–858.

*Jiarui Yao, *Harry Hochheiser, WonJin Yoon, Eli Goldner, and Guergana Savova, editors. 2024. *Overview of the 2024 Shared Task on Chemotherapy Treatment Timeline Extraction, Proceedings of the 6th Clinical Natural Language Processing Workshop, , NAACL June 2024*. Mexico City, Mexico.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. 2021. Biomedical and clinical english model packages for the stanza python nlp library. *Journal of the American Medical Informatics Association*, 28(9):1892–1899.

Yichao Zhou, Yu Yan, Rujun Han, J Harry Caufield, Kai-Wei Chang, Yizhou Sun, Peipei Ping, and Wei Wang. 2021. Clinical temporal relation extraction with probabilistic soft logic regularization and global inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14647–14655.

A Preliminaries

A.1 Pretrained Models

In our experiments, we have employed two main pre-trained model weights available for public use: BART (Lewis et al., 2020), Flan-T5 Chung et al. (2022) which we briefly introduce in this section. The details of how we finetune them for our specific task are described in sections 2.1 and 2.2.

- **BART** (Lewis et al., 2020) BART is a sequence-to-sequence model based on an encoder-decoder architecture, which is composed of a BERT-based bidirectional encoder and an auto-regressive GPT-based left to right decoder. BART is trained by the task of reconstructing a corrupted input sentence into its original text and it has proven to perform well for text-to-text generation tasks such as Summarization.
- **Flan-T5** (Chung et al., 2022) Instruction-tuning is a technique to explicitly guide Large Language models to perform specific tasks. Flan-T5 is a sequence-to-sequence Large Language model that has been fine-tuned using this technique on a mixture of tasks. Flan-T5 has shown performance improvement on unseen tasks.

A.2 LoRA

Large language models inherent to their title have billions of parameters. Finetuning large language models for a specific task or domain is expensive and infeasible in terms of time and computational resource limitations. Hu et al. (2021) introduced Low-Rank Adaptation of Large Language (LoRA) models method to make the finetuning process of these models more efficient and conclusively more accessible by freezing the pre-trained weights of the model and injection of trainable rank decomposition matrices into different layers of the transformer architecture. This method drastically reduces the number of training parameters. It has been shown to perform comparably well to full-parameter finetuning methods and, in some cases, outperforms several baselines with comparable or fewer trainable parameters.

B Detailed Results for the second subtask

We report the detailed results of strict and relaxed settings for all our experiments in the second subtask using the evaluation system in this section. Table 5 contains the results of our experiments for the second subtask. We have experimented with the end2end approach described in section 2.2 using BART and Flan-T5 models with various sizes. Not surprisingly bigger models have performed better across all cancer types for both strict and relaxed evaluation settings. The pipeline approach achieves high recall scores for melanoma and ovarian cancer since it extracts events in a rule-based manner. However, the precision score is low in the pipeline approach, since it identifies drugs and treatments other than chemotherapy-specific ones.

		Micro			Macro Type A			Macro Type B			Official Score
		Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	F1
brca	**Strict**										
	Baseline system	0.622	0.718	0.667	0.811	0.871	0.835	0.663	0.823	0.727	
	Bart-base	0.333	0.256	0.290	0.646	0.640	0.642	0.222	0.207	0.213	
	Bart-large	0.625	0.385	0.476	0.826	0.785	0.797	0.537	0.727	0.455	
	Flan-T5-large + LoRA	0.409	0.231	0.295	0.667	0.645	0.653	0.278	0.220	0.242	
	Flan-T5-xl + LoRA	0.5	0.282	0.360	0.799	0.746	0.763	0.464	0.322	0.369	
	Flan-T5-xxl + LoRA	0.667	0.308	0.421	0.851	0.753	0.781	0.602	0.342	0.417	
	Pipeline approach	0.337	0.718	0.459	0.341	0.451	0.379	0.409	0.702	0.510	
	Relaxed										
	Baseline system	0.795	0.816	0.805	0.866	**0.894**	0.876	0.809	**0.855**	0.837	**0.857**
	Bart-base	0.429	0.353	0.387	0.688	0.668	0.676	0.334	0.281	0.302	0.489
	Bart-large	0.818	0.5	0.621	0.888	0.813	0.837	0.701	0.501	0.564	0.700
	Flan-T5-large + LoRA	0.692	0.529	0.600	0.859	0.769	0.801	0.791	0.552	0.635	0.718
	Flan-T5-xl + LoRA	0.696	0.457	0.552	0.905	0.827	0.853	0.748	0.540	0.607	0.730
	Flan-T5-xxl + LoRA	**0.833**	0.441	0.577	**0.944**	0.830	0.863	**0.851**	0.547	0.634	**0.749**
	Pipeline approach	0.405	**0.833**	0.545	0.381	0.507	0.425	0.515	0.852	0.633	0.529
mela	**Strict**										
	Baseline system	0.667	0.667	0.667	0.571	0.571	0.571	0.357	0.357	0.357	
	Bart-base	0.483	0.333	0.395	0.217	0.119	0.154	0.326	0.179	0.231	
	Bart-large	0.585	0.533	0.558	0.660	0.833	0.658	0.490	0.75	0.487	
	Flan-T5-large + LoRA	0.72	0.4	0.514	0.725	0.682	0.656	0.588	0.524	0.484	
	Flan-T5-xl + LoRA	0.629	0.489	0.550	0.702	0.817	0.714	0.553	0.726	0.571	
	Flan-T5-xxl + LoRA	0.686	0.533	0.6	0.726	0.833	0.733	0.590	0.75	0.6	
	Pipeline approach	0.347	0.911	0.503	0.499	0.865	0.574	0.249	0.798	0.362	
	Relaxed										
	Baseline system	0.630	0.630	0.630	0.570	0.56	0.565	0.354	0.34	0.347	0.456
	Bart-base	0.44	0.458	0.449	0.204	0.167	0.183	0.305	0.25	0.275	0.229
	Bart-large	0.586	0.739	0.654	0.663	0.905	0.694	0.495	0.857	0.542	0.618
	Flan-T5-large + LoRA	0.72	0.565	0.634	0.708	0.690	0.664	0.561	0.536	0.496	0.580
	Flan-T5-xl + Lora	0.667	0.75	0.706	0.698	0.910	0.748	0.548	0.864	0.622	0.685
	Flan-T5-xxl + Lora	**0.731**	0.827	0.775	**0.728**	0.936	0.776	**0.592**	0.905	0.665	**0.720**
	Pipeline approach	0.3375	**1.0**	0.505	0.517	**1.0**	0.608	0.275	**1.0**	0.413	0.511
ovca	**Strict**										
	Baseline system	0.4	0.306	0.347	0.224	0.358	0.239	0.224	0.358	0.239	
	Bart-base	0.350	0.4	0.374	0.391	0.486	0.378	0.391	0.486	0.378	
	Bart-large	0.340	0.423	0.377	0.351	0.357	0.341	0.351	0.357	0.341	
	Flan-T5-large + LoRA	0.494	0.494	0.494	0.471	0.426	0.437	0.471	0.426	0.437	
	Flan-T5-xl + LoRA	0.557	0.518	0.537	0.488	0.559	0.483	0.488	0.558	0.483	
	Flan-T5-xxl + LoRA	0.564	0.411	0.476	0.581	0.545	0.504	0.581	0.544	0.504	
	Pipeline approach	0.265	0.659	0.378	0.297	0.692	0.389	0.297	0.692	0.389	
	Relaxed										
	Baseline system	0.558	0.426	0.483	0.280	0.465	0.329	0.280	0.465	0.329	0.329
	Bart-base	0.434	0.554	0.486	0.440	0.574	0.457	0.440	0.574	0.457	0.457
	Bart-large	0.506	0.620	0.557	0.498	0.590	0.496	0.498	0.590	0.496	0.496
	Flan-T5-large + LoRA	0.633	0.769	0.694	0.581	0.646	0.592	0.581	0.646	0.592	0.592
	Flan-T5-xl + LoRA	0.677	0.646	0.661	0.658	0.677	0.642	0.658	0.677	0.642	0.642
	Flan-T5-xxl + LoRA	**0.686**	0.515	0.588	**0.726**	0.592	0.647	**0.756**	0.592	0.647	**0.647**
	Pipeline approach	0.318	**0.742**	0.445	0.365	**0.812**	0.470	0.365	**0.812**	0.470	0.470

Table 5: System results for the second subtask on the development set