Clinical Language Understanding

2024 Shared Task on Chemotherapy Treatment Timeline Extraction

Xiran Hu, Hiu Lam Choy, Mengtong Guo, Zhaokun Wang, Raziye Sari

Part 2

Task recap

Objective:

- Extract structured chemotherapy timelines (events, temporal expressions, relations) from unstructured clinical notes.
- Subtask 1: Temporal relation classification (gold annotations provided).
- Subtask 2: End-to-end timeline extraction (raw text input).

Dataset:

- Labeled: Breast, ovarian, melanoma cancer EHRs (train/dev/test splits).
- Unlabeled: 57k+ patients' notes for pretraining.

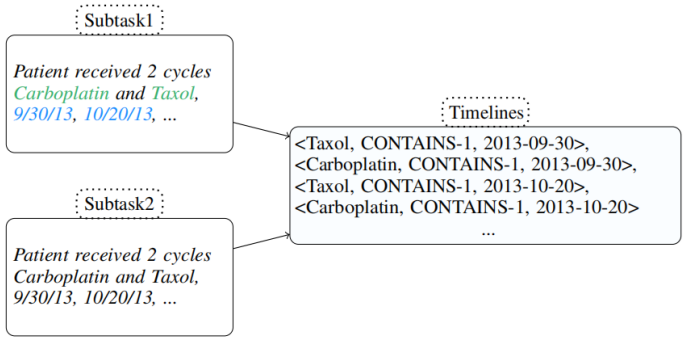


Figure 1: Illustration of the two subtasks in the 2024 Chemotherapy Treatment Timeline Extraction shared task. The input of Subtask1 is patient notes with gold events (highlighted in green) and time expressions (highlighted in blue). The input of Subtask2 is patient notes only. The output of both subtasks is a list of chemotherapy treatment timelines with normalized time expressions. See details in section 2.

Methods recap

Average Scores		Breast Cancer		Melanoma		Ovarian	
Team	Score	Team	Score	Team	Score	Team	Score
LAILab 2	0.70	KCLab 1	0.68	LAILab 2	0.74	LAILab 2	0.74
LAILab 1	0.56	Wonder 2	0.64	LAILab 1	0.57	LAILab 1	0.59
KCLab 1	0.54	Wonder 1	0.63	KCLab 1	0.49	Wonder 3	0.55
Wonder 3	0.53	Wonder 3	0.63	Wonder 3	0.39	Wonder 2	0.55
Wonder 2	0.52	LAILab 2	0.62	Wonder 1	0.39	Wonder 1	0.53
Wonder 1	0.52	LAILab 3	0.53	Wonder 2	0.39	LAILab 3	0.49
LAILab 3	0.47	LAILab 1	0.52	LAILab 3	0.38	KCLAb 1	0.45
NYULangone	0.23	UTSA-NLP 1	0.25	NYULangone	0.32	UTSA-NLP 1	0.19
UTSA-NLP 1	0.22	NYULangone	0.19	UTSA-NLP 1	0.21	NYULangone	0.18
Baseline	0.58	Baseline	0.59	Baseline	0.43	Baseline	0.71

Team	Core Method	Model/Tool	Subtask 1 F1	Subtask 2 F1
NLPeers	Hybrid (Fine-tuned DeBERTa + LLM prompting)	DeBERTa-v3-base, Mixtral	0.77 / 0.64	-
Lexicans	Zero-shot LLM prompting	Llama2, Mistral	0.71	-
NYULangone	Zero-shot prompting (open LLM)	Mixtral 8x7B	-	0.23
KCLab (Today)	End-to-end system	PubMedBERT pipeline	-	0.54
LAILab (Today)	Fine-tuned Seq2Seq models	Flan-T5-xxl, BART-large	**0.90**	**0.70**

- **Fine-tuning >** LLMs: LAILab's small models outperformed large LLMs (Lexicans, NYULangone).
- **Subtask 2 Gap:** End-to-end extraction remains challenging (LAILab leads but scores drop 0.2 vs. Subtask 1).

KCLab & LAILab &UTSA-NLP

KCLab

- **Paper:** KCLab at Chemotimelines 2024: End-to-End System for Chemotherapy Timeline Extraction
- **Authors:** Yukun Tan, Merve Dede, Ken Chen
- **Affiliation:** Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center

Approach:

- UMLS integration.
- Preprocessing and directional filtering.

System Overview

Components:

- Apache cTAKES for chemotherapy terms (e.g., "paclitaxel").
- CLU Lab Timenorm for parsing dates (e.g., "Jun 2008" → 2008-06).
- PubMedBERT for temporal relationship classification.

Enhancements:

- Integration with UMLS for improved term recognition.
- Preprocessing clinical notes to reduce false positives.
- Directional filters for time mention prioritization.

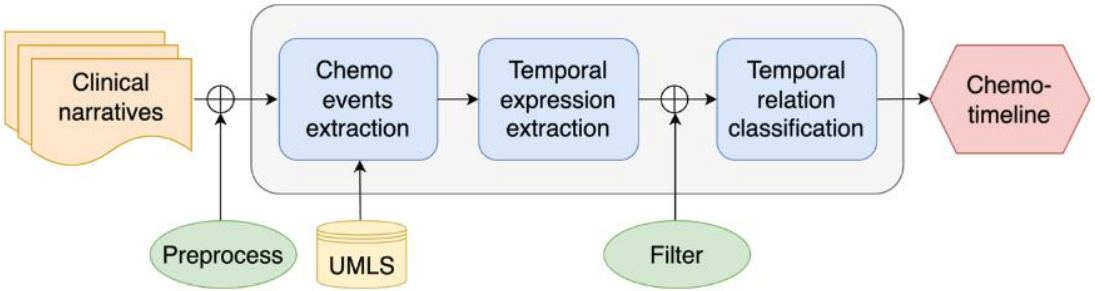
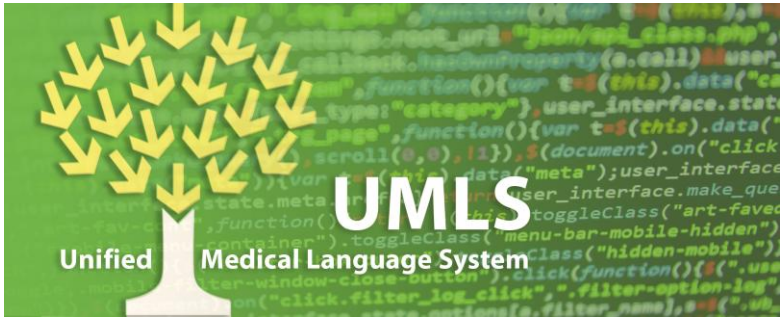


Figure 1: System Overview - Baseline framework enhanced with clinical notes preprocessing, directional time mention filtering, and UMLS integration to extend the extraction dictionary.

1.UMLS(Unified Medical Language System) Integration

- **Objective**: Expand chemotherapy terminology coverage.
- **Method:**Recognizes drug names, synonyms, treatment protocols, and brand names.
- **Impact**:Improves recall and accuracy in extracting chemotherapy-related entities.



- **Example 1**:
  - **Before UMLS**: cTAKES misses "bevacizumab" (anti-angiogenic drug used in chemo combinations).
  - **After UMLS**: Added "bevacizumab" and its synonyms (e.g., "Avastin") → system now recognizes it as a chemo-related agent.
- **Example 2**:
  - **Before UMLS**: "FOLFOX" (a chemotherapy regimen) is not recognized.
  - **After UMLS**: UMLS includes "FOLFOX" and its components (5-FU, oxaliplatin) → system extracts it as a treatment protocol.

2.Clinical Notes Preprocessing

Filtering Non-Relevant Notes:

- **Remove “RAD” and “SP” files (radiation-related or redundant chemotherapy history).**
- **Focus on “NOTE” and “PGN” files with more precise chemotherapy details.**
- **Excluded File**: Patient123_RAD.txt (radiation therapy note) → contains only historical chemo mentions (e.g., "prior chemo in 2019").
- **Retained File**: Patient456_NOTE.txt → includes detailed chemo administration (e.g., "paclitaxel started on 2024-01-01").

Eliminating Redundant Information:

- **Remove file-ending timestamps (already present at the beginning).**
- **Avoid abbreviation conflicts with UMLS terms.**
- **Before**:[Note Footer]: "Documented by Dr. Smith on 2024-06-01 at Houston Clinic."
- **After**: Footer removed to avoid conflicting timestamps.

Fuzzy Recognition for Treatment Plans:

- **Exclude incomplete treatment plans (since they are confirmed in later notes).**
- **Reduces false positives and enhances precision.**
- **Excluded Sentence**: "Plan: Start adjuvant chemo (AC regimen) next Monday if blood counts improve."
- **Reason**: Future plan (not yet confirmed in subsequent notes).

3.Directional Time Mention Filtering

Key Strategy: When multiple time mentions exist in a sentence, **prioritize those appearing after the chemotherapy mention**.

Impact:Reduces temporal misclassification errors.

Case 1 (Same Sentence):
"Resection in 2008; last chemo administered in Nov 2010."

- **Chemo Event**: "chemo"
- **Time Expressions**: "2008" (pre-event), "Nov 2010" (post-event).
- **Filter**: Select "Nov 2010" and discard "2008".

Case 2 (Cross-Sentence):
"Patient completed radiation last week. Today, she received cycle 2 of paclitaxel."

- **Chemo Event**: "paclitaxel"
- **Time Expressions**: "last week" (unrelated sentence), "Today" (same sentence).
- **Filter**: Retain "Today" even though "last week" is post-event but in a different sentence.

Case 3 (Ambiguous Context):
"In 2023, she had chemo; in 2024, she switched to immunotherapy."

- **Chemo Event**: "chemo"
- **Time Expressions**: “2023” (pre-event), “2024” (post-event).
- **Filter: Link "chemo" to "2023" (event time), despite "2024" being later.**

Results Overview

- **F1 Scores:**
  - Breast Cancer: 0.68 (Rank #1)
  - Melanoma: 0.49 (Rank #3)
  - Ovarian Cancer: 0.45 (Rank #7)
- Average ranking: **#3**
- **Improvements over baseline**:
  - **5–10%** F1 gain for breast cancer and melanoma.
  - No improvement for ovarian cancer due to dataset limitations.
- **Breast Cancer**: Largest dataset → robust performance.
- **Ovarian Cancer**: Small dataset + aggressive preprocessing → significant performance drop.
- Improved recall due to UMLS integration.
- False positives from UMLS synonyms (e.g., "vegf trap" vs. "aflibercept").
- Preprocessing & filtering reduced false positives
- Missed true pairs due to filtering steps

Table 3: Final evaluation of test set

Average Scores		Breast Cancer		Melanoma		Ovarian	
Team	Score	Team	Score	Team	Score	Team	Score
LAILab 2	0.70	KCLab 1	0.68	LAILab 2	0.74	LAILab 2	0.74
LAILab 1	0.56	Wonder 2	0.64	LAILab 1	0.57	LAILab 1	0.59
KCLab 1	0.54	Wonder 1	0.63	KCLab 1	0.49	Wonder 3	0.55
Wonder 3	0.53	Wonder 3	0.63	Wonder 3	0.39	Wonder 2	0.55
Wonder 2	0.52	LAILab 2	0.62	Wonder 1	0.39	Wonder 1	0.53
Wonder 1	0.52	LAILab 3	0.53	Wonder 2	0.39	LAILab 3	0.49
LAILab 3	0.47	LAILab 1	0.52	LAILab 3	0.38	KCLAb 1	0.45
NYULangone	0.23	UTSA-NLP 1	0.25	NYULangone	0.32	UTSA-NLP 1	0.19
UTSA-NLP 1	0.22	NYULangone	0.19	UTSA-NLP 1	0.21	NYULangone	0.18
Baseline	0.58	Baseline	0.59	Baseline	0.43	Baseline	0.71

Key Insights from Development Set

- **Preprocessing + UMLS**: Improved precision (0.926 vs. 0.874).
- **Trade-off**: Higher precision but lower recall in Type B due to filtering.

Table 1: Type A evaluation of dev set

		Prec	Recall	F1
Baseline	Breast	0.874	0.894	0.880
	Ovarian	0.648	0.884	0.716
	Melanoma	0.569	0.560	0.565
Proposed	Breast	0.926	0.897	0.909
	Ovarian	0.681	0.851	0.736
	Melanoma	0.570	0.627	0.595

Table 2: Type B evaluation of dev set

		Prec	Recall	F1
Baseline	Breast	0.831	0.885	0.848
	Ovarian	0.648	0.884	0.716
	Melanoma	0.354	0.340	0.347
Proposed	Breast	0.801	0.725	0.757
	Ovarian	0.681	0.851	0.736
	Melanoma	0.355	0.440	0.393

Future Directions

1. **Dictionary Refinement**:
    - Build cancer-type-specific UMLS dictionaries.
    - Create synonym mappings to avoid term duplication.
2. **Data Handling Improvements**:
    - Preserve *"RAD"* and *"SP"* files when no other notes exist.
3. **Model Enhancement**:
    - Explore **ChatGPT** for context-aware TLINK classification (vs. PubMedBERT).
    - Focus on reducing domain-agnostic errors.

Challenges of Rule-Based Approaches:

Challenges

- **Limited Generalization**
  - Hardcoded rules may not adapt well to unseen data or rare chemotherapy terms.
- **Ambiguity Handling**
  - Rules struggle with ambiguous abbreviations and context-dependent meanings.
- **Scalability Issues**
  - Expanding rules for diverse clinical narratives increases complexity and maintenance cost.
- **Precision vs. Recall Tradeoff**
  - Over-filtering can exclude true chemotherapy mentions, while lenient rules may increase false positives.

Improvements

- **Hybrid Approach**
  - Combine rule-based methods with machine learning (LLMs) for better adaptability.
- **Context-Aware Models**
  - Leverage deep learning (e.g., ChatGPT) to improve understanding of medical narratives.
- **Adaptive Filtering**
  - Develop dynamic filtering techniques that adjust based on context and prior knowledge.

LAILab

- **Paper:** LAILab at Chemotimelines 2024: Finetuning sequence-to-sequence language models for temporal relation extraction towards cancer patient undergoing chemotherapy treatment
- **Authors:** Shohreh Haddadan, Tuan-Dung Le, Thanh Duong, Thanh Q. Thieu,
- **Affiliation:** Moffitt Cancer Center and Research Institute, USA,University of South Florida, USA

Approach:

- Text generation
- Instruction-tuning
- Lora

Introduction

- **Objective:** Utilize Flan-T5-xxl for training and apply LoRA for efficient fine-tuning of large models.
- **Focus Areas:**
  - Subtask 1: Reformulating relation classification as a text generation task.
  - Subtask 2: End-to-end vs. Pipeline methods for temporal relation extraction.

- **Flan-T5-xxl:**
  - A powerful instruction-tuned language model.
  - Capable of understanding and generating text based on given instructions.
- **LoRA (Low-Rank Adaptation):**
  - An efficient method for fine-tuning large models.
  - Adds small, trainable rank decomposition matrices to each weight matrix in the model.
  - Reduces computational cost and memory usage while maintaining performance.

LAILab:subtask1

Approach:

- Text generation
- Instruction-tuning
- Lora

Process

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>"

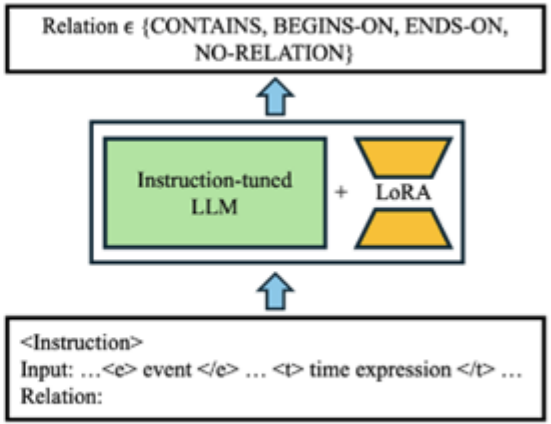


Figure 1: Low-rank adaptation instruction fine-tuning for Subtask 1

Overview of Subtask 1

- **Goal:** Identify temporal relations between chemotherapy events and time expressions in patient EHR notes.
- **Dataset**: Three cancer types (breast, melanoma, ovarian).
- **Dataset Preparation:**
  - Positive instances: Annotated relations from gold standard data.
  - Negative instances: NO-RELATION pairs from EHR notes.
  - To reduce imbalance, NO-RELATION pairs were filtered based on a **250-character distance**

Preprocessing in Subtask 1

- **Sentence Segmentation**: Stanza NLP library("mimic" model).
- **Construction Approaches:**
  - **Concatenated Context**: Two sentences merged if event and time were in different sentences.
  - **Bounded Context**: Included all sentences between event and time expressions.
- **Entity Markers**:Used to distinguish entities in input text.
  - Events: `<e>...</e>`
  - Times: `<t>...</t>`

- **Example:**
  Patient was diagnosed with cancer. <e> Chemotherapy </e> was started. <t> Two weeks ago </t>, treatment was effective.
- ->
- ("Apple Inc.", "founded_by", "Steve Jobs") ("COVID-19", "discovered_in", "2019")

Model for Subtask 1

Model Design

Reformulated as a text generation task:

- Used Large Language Models to **generate** relation types directly.
- Prompt-based conditioning with predefined relation types:
  - CONTAINS, BEGINS-ON, ENDS-ON, NO-RELATION.
- **Models Tested**: Mistral-7B, Flan-T5-xxl, Llama-2-13B.
- **Best Performer**: **Flan-T5-xxl** (instruction-tuned).
- **Task Reformulation**:
  - **Input**: Instruction + context with marked entities.
    "<instruction>Relation Extraction Task: Identify the relation between <e>event</e> and <t>time</t>."
  - **Output**: Directly generate relation type (e.g., "Relation: CONTAINS").
- **LoRA Configuration**:
  - Rank r=16r=16, α=32α=32, applied to Q/K/V/O layers.

Results

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.

	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

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

- Evaluated using timeline score and pairwise temporal classification.
- Metrics: Micro F1 and Relaxed Micro F1 (CONTAINS & BEGINS-ON, CONTAINS & ENDS-ON interchangeable).
- **Best Model:** Flan-T5-xxl (instruction + bounded context) achieved highest scores.
- Bounded context slightly improved relaxed micro F1 compared to concatenated context.
- **Observation:** Classification scores do not correlate well with timeline scores, possibly due to:
  - Macro F1 averaging across all patients.
  - Errors in post-processing (time normalization, deduplication).
- Outperformed baseline in breast & ovarian cancer, matched for melanoma.

Error Analysis

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.

	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

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

Error Analysis:

- **Frequent Errors:**
  - **Spelling mistakes** (e.g., "yesterdat" instead of "yesterday").
  - **Annotation inconsistencies (unlabel & mislabel)** in the dataset.
  - **Complex sentence structures(Tabular data losing structure in plain text format.)** causing misclassification.
- **Key Takeaway:** ENDS-ON relation has the lowest F1 score due to fewer training examples.
- **Future Work:** Enhance low-frequency relation performance (e.g., ENDS-ON) by **data augmentation** and **semi-supervised learning.**

LAILab:subtask2

Approach:

- Text generation
- Instruction-tuning
- Lora

Overview of Subtask 2

- **Goal:** Extract full chemotherapy patient-level timelines from raw EHR notes.
- **Approach 1:** End-to-end **sequence-to-sequence model**
  - Identified **events and time expressions**
  - Classified **temporal relations**
  - Used **Huguet Cabot & Navigli (2021) triplet linearization** to generate target sequences.
- **Approach 2: Pipeline method**
  - **Step 1:** Rule-based extraction of chemotherapy events and time expressions.
  - **Step 2:** Best model from Subtask 1 used for relation classification.

Data Preparation for Evaluation

- **Time Normalization:**
  - Used **document time (DOCTIME)** from EHR headers.
  - Applied **Timenorm** library to normalize relative expressions (e.g., “two weeks ago”, “currently”).
- **Post-processing:**
  - **Filtered out problematic time expressions** (e.g., “1842”, “1000”).
  - **Baseline system** used for de-duplication & final timeline creation.

Approach 1: Seq2Seq Model Architecture

- **How it Works:**
  - **Input:** Raw EHR text
  - **Output:** Directly generates structured triplets.
- **Pretrained Language Model:**
  - **Flan-T5-xxl (11B parameters)**
  - **Instruction-tuned for better generalization**
- **Alternative Models Tested:**
  - BART-large, Mistral-7B, Llama-2-13B-chat
  - **Flan-T5-xxl outperformed all models**
- **Advantages:**
  - Eliminates the need for separate entity extraction & classification.
  - Can capture complex dependencies within text.

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.

Approach 1: Seq2Seq Model Training & Fine-tuning

- **Parameter-Efficient Fine-Tuning (LoRA)**
  - Reduces computational cost while maintaining performance.
  - Fine-tuned on **5 epochs** with **early stopping**.
- **Hyperparameters:**
  - **Max input length:** 256 tokens
  - **Max output length:** 32 tokens
  - **Batch size:** 8
  - **Learning rate:** 3e-5

Approach 2: Pipeline Approach

- **Step 1: Entity Recognition**
  - **Time Expressions:** SUTime (Stanford NLP library).
  - **Chemotherapy Events:** Rule-based matching + Cancer Research UK drug list.
- **Step 2: Relation Classification**
  - Uses **Flan-T5-xxl** (best performer in Subtask 1) for event-time relation classification.
- **Advantages:**
  - More interpretable and computationally efficient.
  - Can be fine-tuned for different datasets.

Approach 2: Step 1 - Entity Extraction

- **Goal:** Identify chemotherapy events and time expressions.
- **Methods:**
  - **Time Expressions:** SUTime (Stanford NLP)
  - **Chemotherapy Events:**
    - Rule-based & dictionary matching
    - Cancer Research UK drug list
    - Stanza NER model for additional recall
- **Example:**
  - **Input:** "Patient started chemotherapy on September 5."

  Output:
  Event: chemotherapy Time: September 5

Approach 2: - Temporal Relation Classification

- **Goal:** Classify relationships between extracted entities.
- **Pre-trained Model:** Flan-T5-xxl (Instruction fine-tuned)
- **Relation Types:**
    - **CONTAINS** - Event happens within time range
    - **BEGINS-ON** - Event starts on the given date
    - **ENDS-ON** - Event ends on the given date
    - **NO-RELATION** - No temporal link
- **Input Format:**
  <instruction> Identify the relation between <e> chemotherapy </e> and <t> September 5 </t>
    - **Model Output:** "BEGINS-ON"

Results

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.

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

- **Best Model:** End-to-end Flan-T5-xxl + LoRA achieved highest results overall.
- Outperformed baseline for melanoma & ovarian cancer but not for breast cancer.
- **Relaxed Setting:** Flan-T5-xxl + LoRA had highest precision across all cancers.
- Rule-based/dictionary-based methods (baseline, pipeline) had higher recall.
- **Limitation:** Poor performance in strict setting due to failure in identifying ENDS-ON relations.

Error Analysis

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.

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

- **Incorrectly Identified Events:**
  - Non-chemotherapy events (e.g., "radiation", "bolus", "augmentin") were mistakenly classified as chemotherapy events.
  - Solution: Keeping all negative instances in training improved filtering of non-chemo events.
- **Dataset Imbalance:**
  - ENDS-ON relation type is underrepresented in melanoma (2%) and ovarian cancer (14%), affecting model accuracy.
- **Unseen Chemotherapy Events:**
  - Some chemotherapy events (e.g., "docetaxel" in test set) were missing in training data.
  - Potential fix: Further refining annotation guidelines.
- **Normalization Errors:**
  - The timnorm library incorrectly resolved two-digit years to the 1900s.
  - Solution: Manual correction or improved time normalization methods.
- **Future Improvements:**
  - Enhance model training with **data augmentation**.
  - Improve **time normalization methods** for better accuracy.

Conclusion & Future Work

- **Summary:**
  - **Seq2Seq excels in accuracy** but requires more computing power.
  - **Pipeline Approach is scalable and interpretable** but slightly less accurate.

- **Next Steps:**
  - Improve **low-frequency relation detection** (e.g., ENDS-ON).
  - Explore **semi-supervised learning** to enhance model performance.
  - **Augmenting data** for **low-frequency relation types**.
  - **Leveraging unlabeled data** to continue **pre-training LLMs**.

Instruction-Tuned and Advanced Approaches

UTSA-NLP

- **Paper:** UTSA-NLP - ChemoTimelines 2024: Evaluating Instruction-Tuned Language Models for Temporal Relation Extraction
- **Authors:** Xingmeng Zhao and Anthony Rios
- **Affiliation:** Department of Information Systems and Cyber Security
  The University of Texas at San Antonio

Approach:

- Instruction-based fine-tuning
- Continued learning

Introduction

- Models fine-tuned for named entity recognition (NER) & relation extraction (RE) on in-domain data often struggle on out-of-domain data
  - Recent zero-/few-shot learning models (CoT-ER, PromptNER, GPT-RE) outperform standard fine-tuned models by leveraging knowledge through prompting
  - → Instruction-based training methods gain significance

Methods

- Instruction tuning
  a. Task-specific QA pairs (NER and RE)
- *Joint* instruction tuning
  a. QA pairs with EHR documents
  b. Jiang et al. 2024 found that placing documents after QA pairs leveraged better concept-specific knowledge access
- *Continued* learning
  a. Unlabeled medical data
  b. Jiang et al. 2024 point to catastrophic forgetting when continuously pre-training on unlabeled data

Methods

1. Instruction tuning OpenChat-3.5-7B on pre-defined relation classes R (EVENTS, DocTimeRel, TIMEX3)
   a. **TLINK temporal relation classification** to determine relation $y \in R \text{ between } (e_{event}, e_{timex3})$
      - Input includes instruction containing context-level knowledge
      - Prompted to output 3-step-reasoning with according text spans
   b. NER
      - Chain-of-thought prompting for entitiy classification
      - Prompted to gather contextual information from *k* sentences before and after

Methods

TLINK temporal relation classification for $y \in R$ between (e_{event}, e_{timex3})

- Input includes instruction containing context-level knowledge

RE System: Please solve the TLINK Relation Extraction task, which is a temporal link only between an EVENT and a TIMEX3. Given the context below, consider what is the most precise relation between two entities belongs to the following N possible relations. The relation to choose must be in these N possible relations: CONTAINS, CONTAINS-1, BEGINS-ON, ENDS-ON. Please output NULL if the EVENT and TIMEX3 entities do not appear in the same sentence or if none of these relations apply.
Question: "Avastin for 6 cycles through March 2009." Given the context, what are the relations between the EVENT entity '{source_entity}' and the TIMEX3 entity '{target_entity}'?
Answer: 1. EVENT entity 'Avastin' is a chemotherapy drug mentioned in the context as part of a treatment regimen that includes carboplatin and paclitaxel, indicating its use in the patient's cancer therapy.
2. TIMEX3 entity 'March 2009' is a time expression indicating the endpoint of the period during which the chemotherapy treatment regimen, including 'Avastin', was administered for 6 cycles.
3. According to the context, the phrase "Avastin for 6 cycles through March 2009" suggests that the chemotherapy treatment phase encompassing 'Avastin' spans until 'April 2013', denoting the EVENT 'Avastin' contains the TIMEX3 'March 2009' within its duration.
Therefore, Avastin | EVENT | March 2009 | TIMEX3 | CONTAINS

Methods

NER

- Input includes instruction containing context-level knowledge

NER System: Given the context below, identify a list of possible entities and for each item explain why it is considered as an entity or not. The response should be structured as follows: 'entity name | entity type | True/False | Explanation', where you explain the rationale behind the classification. Output NULL and mark it as False if there is no entity identified.
Define: the DOCTIME entity refers to the time expression representing the document creation time, usually found at the start of the document.
Question: "{DOCTIME}" Given the context, the DOCTIME entity is:
Answer: 20090824 | DOCTIME | True | As it is listed as the "Principal Date" at the start of the document, indicating it as the date the document was created or formalized.

Define: The EVENT entity refers to chemotherapy mention in the clinical notes, including general terms like 'chemotherapy' and 'chemo', as well as specific chemotherapy treatments such as 'cytoxan', which involve the use of powerful drugs to target and destroy cancer cells, often administered in cycles to shrink tumors, prevent cancer spread, and potentially achieve remission or alleviate symptoms. Diseases (e.g., "melanoma"), diagnostic scans (e.g., "FDG PET scan," "CT scan") or medications not used in chemotherapy (e.g., "Vicodin" for pain relief, "Zocor" for cholesterol management) are not EVENT entities.
Question: "Avastin for 6 cycles through March 2009." Given the context, all relevant EVENT entities are:
Answer: Avastin | EVENT | True | As it is a specific type of chemotherapy treatment for breast cancer, the mention of Avastin highlights a particular therapeutic approach within the patient's care.

Experiments

- Data
  - EHR documents from the University of Pittsburgh/UPMC:
  - 62,000 unlabeled patient documents on breast/ovarian cancer & 16,000 on melanoma cancer
  - 310 gold-annotated patients' histories
    - EVENT: Any relation to document creation time `(BEFORE, BEFORE-OVERLAP, OVERLAP, AFTER)`
      TIME: Using TimeNorm (Laparra et al., 2018; Xu et al., 2019).
      Temporal relation TLINKs: Link EVENT & TIMEx3 `(CONTAINS, CONTAINS-1, BEFORE, BEGINS-ON, ENDS-ON)`
- Training
  - Low-Rank Adaptation to optimize specific target modules and computing average negative log-likelihood loss
  - For QA+doc: next token prediction loss on the document's tokens
  - Best settings: temperature: 0.2, top p: 0.5 and top k: 20

Experiments

- Metrics
  - Subtask 1: Using gold-standard DOCTIME annotations
    - Subtask 2: Flter out those without DOCTIME prediction
  - Normalize time expressions and filter out duplicate time-event pairs.
  - F1-score is computed for each patient of (“chemo EVENT”, “temporal relation”, “TIMEX3”) tuples and averaged over all patients to obtain macro F1 score

Results

	Breast			Melanoma			Ovarian			Total Average
	Type A	Type B	Average	Type A	Type B	Average	Type A	Type B	Average	
train QA	.81	.50	.66	.80	.70	.75	.57	.57	.57	.66
+ train QA + DOC	.82	.51	.67	.83	.74	.78	.58	.58	.58	.68
+ train on unlabeled corpus	.77	.39	.58	.80	.70	.75	.56	.56	.56	.63

Table 1: Official results on the dev set for subtask 1.

- Best performance achieved through QA pairs with associated documents
  - Timeline relation extration with total average precision of **.68**

Team	Breast	Melanoma	Ovarian	Total Average
LAILab 1	.96	.87	.88	.90
Wonder 2	.90	.84	.77	.84
NLPeers 1	.72	.81	.75	.77
BioCom 1	.88	.61	.72	.74
Lexicans 1	.68	.83	.61	.71
UTSA-NLP 1 (Ours)	.70	.68	.69	.69
EmoryClincalRXMiners 1	.44	.47	.34	.40
Baseline	.93	.87	.88	.89

Table 3: Official results on the test set for subtask 1.

Results

	Breast			Melanoma			Ovarian			Total Average
	Type A	Type B	Average	Type A	Type B	Average	Type A	Type B	Average	
train QA + Doc	.78	.41	.59	.71	.57	.64	.17	.17	.17	.47

Table 2: Official results on the dev set for subtask 2.

- Subtask 2 performed worse
  - NER struggles with ovarian cancer type (**.17** accuracy), with total avg. precision of .47 → complex cancer type

Model	Breast	Melanoma	Ovarian	Total Average
LAILab 2	.62	.74	.74	.70
KCLab 1	.68	.49	.45	.54
Wonder 3	.63	.39	.55	.53
NYULangone	.19	.32	.18	.23
UTSA-NLP (Ours)	.25	.21	.18	.22
Baseline	.59	.43	.71	.58

Table 4: Official results on the test set for subtask 2.

Limitations

- Continued pre-training on unlabeled data decreases performance
  - training on 1% of unlabeled data
- Restricted negative examples for RE QA pairs (3 unrelated)
  - Missing negative examples for NER QA pairs

Conclusion

- Performance on both subtasks lower than EntitiyBERT baseline
  - Subtask 1: High amount of false positives for non-existent event-time negative examples
  - Subtask 2: Misidentification of EVENTS, e.g.: diseases, diagnostic scans/codes, people and non-chemotherapy medications, despite post-processing with RegEx
- Generative models lack high specificity that is required for the NER/RE tasks

Joint Q&A - Outlook

Thank you!