Wonder at Chemotimelines 2024: MedTimeline: An End-to-End NLP System for Timeline Extraction from Clinical Narratives

Liwei Wang, Qiuhao Lu, Rui Li, Sunyang Fu, and Hongfang Liu
School of Biomedical Informatics,
The University of Texas Health Science Center at Houston,
Houston, TX, USA
{liwei.wang, qiuhao.lu, rui.li.1, sunyang.fu, hongfang.liu}@uth.tmc.edu

Abstract

Extracting timeline information from clinical narratives is critical for cancer research and practice using electronic health records (EHRs). In this study, we apply MedTimeline, our end-to-end hybrid NLP system combining large language model, deep learning with knowledge engineering, to the ChemoTimeLine challenge subtasks. Our experiment results in 0.83, 0.90, 0.84, and 0.53, 0.63, 0.39, respectively, for subtask1 and subtask2 in breast, melanoma and ovarian cancer.

1 Introduction

Patients' medical history plays a crucial role in guiding the decisions made by clinicians. Yet, the vast majority of temporal information, along with the medical events, is embedded in clinical narratives. For instance, details such as the timing of chemotherapy administration for cancer patients, particularly those referred to the current hospital from other healthcare facilities, are often documented within clinical notes during patient consultations with physicians. There is a pressing need to automatically extract timeline information from clinical narratives to facilitate the understanding of disease progression and treatment efficacy and enhance the quality of cancer research and patient care based on electronic health records (EHRs). Large language models (LLMs), trained on a large amount of unstructured text and then applied to a task through instructive prompts (Tam et al., 2023), have recently shown great value in information extraction and garnered significant attention. We developed MedTimeline, an end-to-end hybrid natural language processing (NLP) system, which combines LLMs and deep learning to support knowledge engineering for timeline information extraction. In this ChemoTimeLine challenge, we applied MedTimeline to the two subtasks and had it evaluated based on the tasks-specific data(Yao et al., 2024).

2 Related Work

In the 2012 i2b2 clinical temporal relations challenge, Sohn *et al.* constructed an automated system, i.e., MedTime, that leveraged the framework of HeidelTime, for TIMEX3 extraction from clinical text (Sohn et al., 2013). The system extracts temporal information, including date, time, duration, and frequency, along with their normalized values, demonstrating superior performance. In addition, using the THYME corpus (Styler IV et al., 2014), Liu *et al.* developed an attention-based neural network model to extract containment relations within sentences of clinical narratives (Liu et al., 2019), which outperformed the existing state-of-the-art neural network models at the time.

NLP systems derived from challenges are usually limited to functioning within the confines of the tasks they're specifically designed for. Consequently, Wang *et al.* further expanded their NLP work to patient-level event temporal relation extraction based on real EHR data (Wang et al., 2019). Their results revealed that complete data related to patients' journeys was important for accurate identification of diagnosis dates. In addition, domain knowledge, e.g., chemotherapy drug and transplant names of multiple myeloma, and histology cell type of lung cancer were critical for event temporal relation extraction. In addition, this study demonstrated the usability of MedTime and MedTagger, resource-driven open-source UIMA-based frameworks with the capacity to incorporate knowledge engineering (Sohn et al., 2013; Liu et al., 2013; Wen et al., 2019), for EHR-based cancer research.

3 Methodology

In this section, we present our solution, MedTimeline, an end-to-end NLP system comprising an event entity (Chemotherapy entity for subtask 2) extractor, a temporal entity extractor (subtask 2), and a patient-level timeline aggregator (subtasks 1

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 483–487

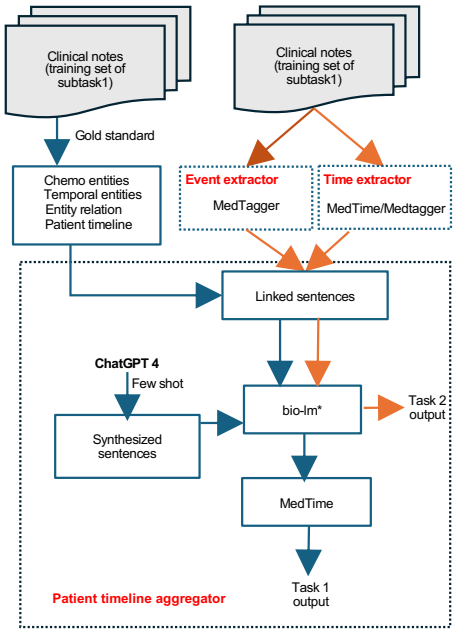

Figure 1: Architecture of MedTimeline

and 2). The architecture of MedTimeline includes two well-established knowledge engineering NLP pipelines (MedTagger as event extractor and MedTime as temporal expression extractor) from the Open Health NLP (OHNLP) consortium, a context-aware deep learning open-source architecture, and an LLM-empowered data augmentation pipeline (Figure 1). Specifically, the data augmentation pipeline incorporates ChatGPT to generate synthetic data to facilitate the fine-tuning of a pre-trained language model for temporal relation classification within the timeline aggregator.

3.1 Event Entity Extractor

MedTimeline leverages MedTagger for event entity extraction. Particularly, the knowledge artifacts of chemotherapy drug names for breast cancer, ovarian cancer and melanoma, are first collected from both the training data set and the online knowledge hub of the American Cancer Society[1], and then made into a MedTimeline rule set that is compatible with MedTagger.

3.2 Temporal Entity Extractor

MedTime and MedTagger function as the temporal entity extractors in the MedTimeline to automatically extract temporal information from clinical notes. For MedTime, missing temporal expression rules are added to MedTime through the comparison of the results automatically extracted by MedTime (existing rules) with the gold standards of the training set, i.e., subtask1 in this study. For instance, we add "at this time" and "on the day" rules into MedTime. Additionally, we leverage MedTagger to manage complex rules that can not be added to MedTime, in order to extract the temporal information not captured by MedTime. For example, MedTime failed to extract "today" when it was preceded by a number, e.g., "5 today". We made a regular expression rule for this case to enable automatic extraction of "today".

[1] https://www.cancer.org/

w/o Synthetic Data			w/ Synthetic Data		
Relation	Train	Dev	Relation	Train	Dev
Breast Cancer			**Breast Cancer**		
OPEN	389	133	OPEN	389	133
CONTAINS	298	57	CONTAINS	492	57
BEGINS-ON	131	27	BEGINS-ON	231	27
ENDS-ON	26	29	ENDS-ON	225	29
Melanoma			**Melanoma**		
OPEN	35	192	OPEN	35	192
CONTAINS	37	157	CONTAINS	37	157
BEGINS-ON	10	42	BEGINS-ON	205	42
ENDS-ON	1	2	ENDS-ON	191	2
Ovarian Cancer			**Ovarian Cancer**		
OPEN	338	226	OPEN	338	226
CONTAINS	327	140	CONTAINS	516	140
BEGINS-ON	98	34	BEGINS-ON	266	34
ENDS-ON	59	52	ENDS-ON	256	52

Table 1: Dataset statistics with and without synthetic data.

3.3 Synthetic Data Augmentation

Training data insufficiency and imbalance are critical issues as they may impact the quality and reliability of predictive models (Lu et al., 2021). To address these issues, MedTimeline synthesizes artificial data to enrich the training data and facilitate model training. Essentially, ChatGPT-4 (i.e., `gpt-4-1106-preview`) prompting is used to generate synthetic data.

In the context of the challenge subtasks, we identified the lack of sufficient data for such a condition as melanoma, and imbalance of the datasets concerning the three temporal relations during the initial data analysis. We instruct ChatGPT-4 to produce artificial data, as shown in Table 1. Specifically, textual segments extracted between chemotherapy events and time expressions demonstrate a unique pattern for each predefined tem-

poral relation as well as each cancer type, e.g., BEGINS-ON of melanoma is substantially different from ENDS-ON of breast cancer. Following the patterns, we manually design 5 example text pieces for each temporal relation of each cancer type to use as few-shot demonstrations. Notably, we only synthesize textual segments connecting chemo and time instead of the entire clinical note, and their numbers are determined based on preliminary experiments. We use the following prompt:

You are a helpful assistant in synthetic data generation. Your job is to generate a sentence containing a chemotherapy entity for melanoma (source) and a TIMEX3 entity (target). The relation between them is ENDS-ON. After reading and comprehending the examples, generate 50 data samples. The outputs should be in three columns: source, target and context. Use | as the delimiter and do not add index numbers to the generated samples. Be diverse, representative, and accurate, e.g., the chemo should be for the specific cancer and do not mention the specific cancer in the sentence. Examples: [manually-designed 5-shot demonstrations] Generated Data:

3.4 Relation Extraction

We cast relation extraction for the medical events and temporal expressions as a multi-class text classification problem. Essentially, we extract the textual segment (e.g., *"Chemo started Today."*) that links a chemotherapy event (e.g., *chemo*) with its related time expression (e.g., *today*) from the clinical note. We then categorize the textual segment into one of the predefined temporal relations.

The problem is also an open-world classification problem (Bai et al., 2022), as it requires the model to predict a sample as OPEN, which indicates it has an open/unspecified temporal relation or does not have any relation. To create the corresponding training data for this category, we adopt a simple yet effective negative sampling strategy where we extract <*chemo, time*> pairs in the training set whose distance is less than 250 characters[2] and do not belong to any of the predefined temporal relations. We consider such negative samples to be hard and realistic. It is worth noting that since candidate relations are not provided in the test set, we use the same strategy for candidate search during inference.

Formally, given a clinical note S containing a list of chemotherapy events $\mathcal{C} = \{c_1, c_2, \ldots, c_{|\mathcal{C}|}\}$ and a list of time expressions $\mathcal{T} = \{t_1, t_2, \ldots, t_{|\mathcal{T}|}\}$, we search for candidate pairs using the aforementioned strategy and extract the text between them as input $\mathcal{D} = \{x_i, x_2, \ldots, x_k, \ldots, x_{|\mathcal{D}|}\}$ where x_k is the text between $c_i \in \mathcal{C}$ and $t_j \in \mathcal{T}$. The objective is to predict the corresponding label $y_k \in \mathcal{E}$ where $\mathcal{E} = \{\texttt{CONTAINS-1}, \texttt{BEGINS-ON}, \texttt{ENDS-ON}, \texttt{OPEN}\}$.

In particular, we use the bio-lm[3] pre-trained language model (Lewis et al., 2020) to encode the text and feed the representation for the [CLS] token in the last layer into a linear layer for classification. The model is optimized with cross-entropy loss:

$$\mathcal{L} = -\sum_{l=1}^{4} y_l \log \hat{y}_l \quad (1)$$

where y_l is the ground-truth label and $\hat{y}_l$ refers to the output prediction probabilities.

3.5 Time Expression Normalization

We adapt MedTime[4] to convert temporal expression from clinical notes into standardized TIMEX3 format. Types of MedTime output include standard dates and time intervals. For time entities which are directly mapped into standard dates such as *2013-11-12* and *2012-W06*, the MedTime output is used as the standardized TIMEX3 date. For time entities which are mapped into time intervals, the standardized TIMEX3 date is calculated by subtracting time intervals from the principal date.

3.6 Patient-level Timeline Aggregation

If the relation of the pair is classified as OPEN by bio-lm, we do not assign any specific temporal relation for the pair. We then employ the aforementioned temporal expression normalization method to convert the temporal entities into standardized TIMEX3 format. At last, we aggregate all <*chemo, time*> pairs whose relation are not OPEN to construct the patient-level timeline.

4 Experiments

4.1 Results

In this section, we first show the statistics of the dataset with and without synthetic data in Table 1. We then present the temporal relation classification results across different models and cancers in Table 2. Finally, we show the patient-level timeline extraction results on the dev and test sets, as shown

[2]Maximum distance among <*chemo, time*> pairs of a predefined temporal relation in the training set.

[3]We use the best-performing variant RoBERTa-large-PM-M3-Voc across all experiments in this work.

[4]https://github.com/OHNLP/MedTime

Cancers	Models	CONTAINS			BEGINS-ON			ENDS-ON			OPEN			Overall			
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	Acc	P	R	F1
Breast	PubMedBERT-tlink	0.829	0.509	0.630	0.556	0.741	0.635	0.833	0.345	0.488	0.912	0.857	0.884	0.703	0.844	0.703	0.751
	BioClinicalBERT	0.636	0.860	0.731	0.870	0.741	0.800	0.727	0.276	0.400	0.956	0.970	0.963	0.837	0.845	0.837	0.825
	BioClinicalBERT*	0.831	0.860	0.845	0.828	0.889	0.857	0.917	0.759	0.830	0.948	0.955	0.951	0.902	0.904	0.902	0.902
	bio-lm	0.773	0.895	0.829	0.920	0.852	0.885	0.769	0.345	0.476	0.901	0.962	0.931	0.862	0.858	0.862	0.849
	bio-lm*	0.817	0.860	0.838	0.897	0.963	0.929	0.909	0.690	0.784	0.978	0.993	0.985	0.923	0.923	0.923	0.921
Melanoma	PubMedBERT-tlink	0.617	0.586	0.601	0.914	0.762	0.831	0.000	0.000	0.000	0.730	0.745	0.737	0.679	0.701	0.679	0.689
	BioClinicalBERT	0.479	0.994	0.646	0.000	0.000	0.000	0.000	0.000	0.000	0.985	0.344	0.510	0.565	0.672	0.565	0.507
	BioClinicalBERT*	0.580	0.949	0.720	0.035	0.048	0.040	0.000	0.000	0.000	0.948	0.380	0.543	0.570	0.698	0.570	0.557
	bio-lm	0.596	0.968	0.738	0.000	0.000	0.000	0.000	0.000	0.000	0.957	0.688	0.800	0.723	0.705	0.723	0.686
	bio-lm*	0.569	0.949	0.711	0.925	0.881	0.902	0.000	0.000	0.000	0.923	0.438	0.594	0.687	0.777	0.687	0.671
Ovarian	PubMedBERT-tlink	0.807	0.507	0.623	0.392	0.588	0.471	0.435	0.192	0.267	0.942	0.929	0.935	0.688	0.800	0.688	0.727
	BioClinicalBERT	0.750	0.879	0.809	0.615	0.471	0.533	0.895	0.327	0.479	0.918	0.987	0.951	0.839	0.840	0.839	0.821
	BioClinicalBERT*	0.774	0.907	0.836	0.800	0.588	0.678	0.920	0.442	0.597	0.929	0.978	0.953	0.865	0.870	0.865	0.855
	bio-lm	0.703	0.879	0.781	0.667	0.588	0.625	0.905	0.365	0.521	0.951	0.951	0.951	0.834	0.848	0.834	0.824
	bio-lm*	0.778	0.850	0.812	0.800	0.706	0.750	0.906	0.558	0.690	0.920	0.965	0.942	0.863	0.865	0.863	0.858

Table 2: Temporal relation classification performance across different models and cancers with relation-wise and overall scores. * represents fine-tuning with synthetic data.

in Table 3. All experimental results are obtained during the challenge.

We use three pre-trained language models in the clinical domain as baselines, i.e., PubMedBERT-tlink[5], BioClinicalBERT (Alsentzer et al., 2019), and bio-lm (Lewis et al., 2020). Note that we do not fine-tune PubMedBERT-tlink as it is already trained on a similar task and data. For relation classification, we use precision (P), recall (R), and F1-score as the metrics. For patient-level timeline extraction, we use the official script of the challenge where the *relaxed-to-month* F1-score is used as the metric. One key observation is that both BioClinicalBERT and bio-lm demonstrate a significant improvement with synthetic training data, highlighting the effectiveness of data augmentation in this context. All models struggle with `ENDS-ON` for Melanoma even after training data is augmented from 1 to 191. The reason lies in the fact that there are very limited data samples in the dev set, i.e., only 2 samples in the dev set as shown in Table 1.

4.2 Error analysis

We compare the patient-level chemo-timeline generated by MedTimeline with the gold standard of dev set to identify errors from our system. The errors mainly originate from two sources, i.e., time normalization and relation classification. The former is caused by wrong anchor time retrieved from MedTime and inaccurate imputation of the incomplete time entity. The latter arises from incomplete and complex text input. Incomplete text input is caused by our strategy of merely extracting the text between the chemo entity and time entity, leading to the missingness of some useful information. For example, the original text in clinical notes is *'She received her 9th and final dose of IL2 at 9/22'*[6], and the timeline in the gold annotation is *["il2", "ends-on", "2012-09-22"]*. However, by extracting *'IL2 at 9/22'* as input, our system wrongly classifies the relation as `BEGINS-ON`. Meanwhile, some text input is too complex for the system to classify the correct relation. For example, given the original text *'Today he feels well. He had been able to control the symptoms of nausea that he has experienced with his TCH chemotherapy'*, our system wrongly classifies the relation as `ENDS-ON` while the relation should be `OPEN`.

Subtask	Split	Breast	Melanoma	Ovarian
Subtask 1	Dev	0.86	0.80	0.77
	Test	0.83	0.90	0.84
Subtask 2	Dev	0.83	0.71	0.75
	Test	0.53	0.63	0.39

Table 3: Patient-level timeline evaluation results for Subtasks 1 and 2.

5 Conclusion

We present MedTimeline, an end-to-end hybrid NLP system generalizable to any medical events for patients' timeline extraction, and evaluate it based on the ChemoTimeLine challenge data. Our system ranks the second place in subtask 1 and the third place in subtask 2. In the future, we will continue to develop the MedTimeline, and tailor it to the scenarios of various medical events.

[5] https://huggingface.co/HealthNLP/pubmedbert_tlink

[6] All examples are rephrased in order to avoid data leakage

Acknowledgment

This project is supported by the Cancer Prevention Research Institute of Texas (CPRIT) RR230020, National Institute of Aging grant RF1AG072799, National Human Genome Research Institute R01HG12748, and National Library of Medicine R01LM11934.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Ke Bai, Guoyin Wang, Jiwei Li, Sunghyun Park, Sungjin Lee, Puyang Xu, Ricardo Henao, and Lawrence Carin. 2022. Open world classification with adaptive negative samples. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4378–4392, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.

Hongfang Liu, Suzette J Bielinski, Sunghwan Sohn, Sean Murphy, Kavishwar B Wagholikar, Siddhartha R Jonnalagadda, KE Ravikumar, Stephen T Wu, Iftikhar J Kullo, and Christopher G Chute. 2013. An information extraction framework for cohort identification using electronic health records. *AMIA Summits on Translational Science Proceedings*, 2013:149.

Sijia Liu, Liwei Wang, Vipin Chaudhary, and Hongfang Liu. 2019. Attention neural model for temporal relation extraction. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 134–139.

Qiuhao Lu, Dejing Dou, and Thien Huu Nguyen. 2021. Textual data augmentation for patient outcomes prediction. In *2021 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 2817–2821. IEEE.

Sunghwan Sohn, Kavishwar B Wagholikar, Dingcheng Li, Siddhartha R Jonnalagadda, Cui Tao, Ravikumar Komandur Elayavilli, and Hongfang Liu. 2013. Comprehensive temporal information detection from clinical text: medical events, time, and tlink identification. *Journal of the American Medical Informatics Association*, 20(5):836–842.

William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C De Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. 2014. Temporal annotation in the clinical domain. *Transactions of the association for computational linguistics*, 2:143–154.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255.

Liwei Wang, Jason Wampfler, Angela Dispenzieri, Hua Xu, Ping Yang, and Hongfang Liu. 2019. Achievability to extract specific date information for cancer research. In *AMIA Annual Symposium Proceedings*, volume 2019, page 893. American Medical Informatics Association.

Andrew Wen, Sunyang Fu, Sungrim Moon, Mohamed El Wazir, Andrew Rosenbaum, Vinod C Kaggal, Sijia Liu, Sunghwan Sohn, Hongfang Liu, and Jungwei Fan. 2019. Desiderata for delivering nlp to accelerate healthcare ai advancement and a mayo clinic nlp-as-a-service implementation. *NPJ digital medicine*, 2(1):130.

Jiarui Yao, Harry Hochheiser, WonJin Yoon, Eli Goldner, and Guergana Savova. 2024. Overview of the 2024 shared task on chemotherapy treatment timeline extraction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*. NAACL.