KCLab at Chemotimelines 2024: End-to-end system for chemotherapy timeline extraction – Subtask2

Yukun Tan, Merve Dede, Ken Chen

Department of Bioinformatics and Computational Biology,
The University of Texas MD Anderson Cancer Center, Houston, Tx, USA

{ytan1, mdede, kchen3}. mdanderson.org

Abstract

This paper presents our participation in the Chemotimelines 2024 subtask2, focusing on the development of an end-to-end system for chemotherapy timeline extraction. We initially adopt a basic framework from subtask2, utilizing Apache cTAKES for entity recognition and a BERT-based model for classifying the temporal relationship between chemotherapy events and associated times. Subsequently, we enhance this pipeline through two key directions: first, by expanding the exploration of the system, achieved by extending the search dictionary of cTAKES with the UMLS database; second, by reducing false positives through preprocessing of clinical notes and implementing filters to reduce the potential errors from the BERT-based model. To validate the effectiveness of our framework, we conduct extensive experiments using clinical notes from breast, ovarian, and melanoma cancer cases. Our results demonstrate improvements over the previous approach.

1 Introduction

In recent years, the rapid development and widespread implementation of Electronic Health Records (EHRs) have created a significant demand for the clinical notes processing and information extraction within the realm of medical research. (Yanshan Wang et al., 2018) Particularly, the extraction of temporal information, encompassing temporal expressions, temporal events, and temporal relations, has created new opportunities for dynamic treatment studies(Sun et al., 2013; UzZaman et al., 2014). Among the various treatment modalities, chemotherapy stands out as one of the most critical and widely used approaches in cancer therapy. EHRs with temporal information offer a unique advantage by providing a chronological roadmap of patient-specific treatments. These timelines play a key role in understanding the effectiveness of chemotherapy, evaluating treatment responses, and identifying patterns in patient outcomes. They serve as invaluable resources for researchers investigating the interplay among treatment protocols, tumor biology, and patient characteristics. By analyzing these timelines, researchers can uncover trends, predictors of response and potential markers for treatment success or failure. As such, the construction of accurate and comprehensive chemotherapy treatment timelines is not only an academic pursuit but a practical clinical necessity in advancing cancer care and improving outcomes.

However, this task presents notable challenges due to the domain-specific nature of EHRs, namely, variations in writing style and quality, lack of text structure, and the pervasive presence of redundant information. Moreover, the creation of annotated corpora manually is a resource-intensive process, demanding substantial human effort and time. Consequently, numerous research efforts have turned to employ rule-based, machine-learning, or hybrid methods to extract general temporal information from clinical narratives (Moharasan & Ho, 2019; Najafabadipour et al., 2020; Liwei Wang et al., 2020). Notably, despite these efforts, there are currently no available tools designed specifically for extracting timelines to contextualize cancer treatment. Hence, this competition subtask aims to fill this gap by developing an end-to-end system for chemotherapy timeline extraction (*Jiarui Yao et al., 2024). This system not only addresses the urgent need for accurate and

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 417–421
June 21, 2024 ©2024 Association for Computational Linguistics

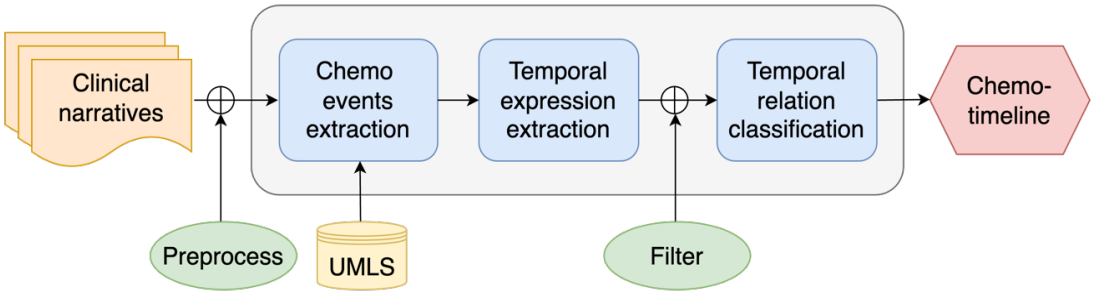

Figure 1: System Overview - Baseline framework enhanced with clinical notes preprocessing, directional time mention filtering, and UMLS integration to extend the extraction dictionary.

comprehensive timelines but also showcases the potential of leveraging advanced computational methods to enhance cancer care practices.

2 System Overview

2.1 Baseline framework

The pipeline mainly combines three main software packages: Apache cTAKES (Savova et al., 2010), CLU Lab Timenorm (Laparra et al., 2018; Xu et al., 2019), and Huggingface transformers. cTAKES, a java-based tool, offers powerful text engineering and information extraction capabilities, particularly tailored for clinical text. It utilizes the cTAKES Python Bridge to Java (ctakes-pbj) to process text artifacts seamlessly in Python, leveraging cTAKES' modules for entity recognition and sentence tokenization. CLU Lab Timenorm is employed for identifying and normalizing date and time expressions. The pipeline has incorporated a customized version of Timenorm into the pipeline, which allows for improved handling of approximate dates, a common occurrence in clinical narratives. This step ensures consistency and standardization of temporal representations. Huggingface Transformers is a widely used deep learning library for natural language processing tasks. This pipeline employs the PubMedBERT – based model (Gu et al., 2022) (Temporal Link – TLINK model) to identify and classify the temporal relationships between chemotherapy mentions and their associated dates. The classifier determines the temporal relationship between each paired mention, whether it's "begin on", "end on", "contain-1", or "none".

2.2 UMLS integration

We have enhanced the capabilities of cTAKES by integrating the Unified Medical Language System (UMLS), thereby extending its dictionary to have a more comprehensive range of chemotherapy terminologies. We evaluated all medically relevant concepts in the UMLS database related to chemotherapy as well as their descendant terms to obtain a complete hierarchy. This integration allows cTAKES to recognize and extract a broader array of chemotherapy-related terms, including generic drug names, their synonyms, treatment protocols, drug brand names and so on. This approach ensures that the system captures a more exhaustive list of chemotherapy-related terms and agents, thus improving the completeness of extracted information.

2.3 Clinical notes preprocessing

In our clinical notes preprocessing stage, we implemented several steps to enhance the efficiency and accuracy of information extraction. We examined the provided notes carefully to evaluate their structures, and to understand the content of information in each note category. After the evaluation period, firstly, we removed files with names ending in "RAD" or "SP", as these notes often did not contain any chemotherapy information or only contained redundant chemotherapy history of the patients, which were already present in other clinical notes. For example, the files with "RAD" may contain the information related to radiation procedures or outcomes, which occasionally included descriptions of chemotherapy in the patient history statement. We determined these sections to be redundant, as more detailed and clearer descriptions were typically already found in files ending with "NOTE" or "PGN". Secondly, we eliminated the concluding portions of files containing information about the person recording the note, time, and location. While these timestamps may initially be perceived as valuable, they are usually redundant as timestamps were typically provided at the beginning of each record. Additionally, these

sections often contained abbreviations that overlapped with abbreviations in our expanded UMLS dictionary for chemotherapy agents, leading to false positives. Thirdly, we employed fuzzy recognition to filter out paragraphs related to treatment plans. Since current treatment plans are incomplete until documented as finished in subsequent notes, we can confidently exclude these sections without missing relevant information. This step effectively reduced the occurrence of false positives, as changes to treatment plans were noted elsewhere in subsequent records. These preprocessing steps not only simplified the data but also significantly enhanced the precision of our information extraction process, ensuring that extracted chemotherapy-related details are accurate and comprehensive.

2.4 Directional time mention filtering

After successfully extracting chemotherapy events and temporal expression pairs, we introduced a novel filter prior to the temporal relation classification step, focusing on the directionality of time mentions. This filter aims to reduce potential errors in classification by considering the ordering of temporal expressions in relation to the chemo events. Specifically, when multiple temporal expressions surround a chemo event within the same sentence and appear after the chemo event, we prioritize these temporal expressions over those occurring before the chemo event. For instance, in the sentence "He had resection in Jun 2008, last chemo was in Nov 2010," we identified the temporal expressions "Jun 2008" and "Nov 2010." In this case, we disregard the time preceding the chemotherapy event since a temporal expression already exists in the same sentence following the chemotherapy event, making it clear that "Nov 2010" pertains to the chemotherapy event. Likewise, in the sentence "She presents to the office on today's date for the chemo as per the standard FDA approved regimen. She also did radiation last week," we detected the temporal expressions "today's date" and "last week" surrounding the chemotherapy event. However, since "last week" is not in the same sentence as the chemotherapy event, we do not ignore the temporal expression "today's date." This analysis of directional cues in time mentions is crucial for our task.

While theoretically, the BERT – based model's classification (TLINK) could address this by categorizing irrelevant times as "none", our findings suggest that due to potential limitations in training data, this classification may not always be accurate, particularly in scenarios where temporal expressions occur both before and after the chemotherapy terminology. Our introduced filter significantly reduces the chances of misclassifications, thereby enhancing the accuracy and robustness of our temporal relation classification system.

3 Results

In Chemotimelines 2024 subtask2, our team achieved the 3rd highest rank in the average scores, with F1 scores of 0.68 for breast cancer (rank #1), 0.49 for melanoma (rank #3), and 0.45 for ovarian cancer (rank #7) (Table 3). These scores were calculated by averaging type A and type B evaluation metrics. Type A includes notes without true relations, while type B excludes such notes. Comparing our results to baseline performance, we observed an improvement of around 5%-10% for breast cancer and melanoma, while not for ovarian cancer.

Due to the unavailability of the test set, we present results from the development set to analyze the strengths and weaknesses of our pipeline. As depicted in Table 1 (Type A) and Table 2 (Type B), our system generally outperforms the baseline in terms of recall, attributed to the integration of the UMLS dictionary. However, this integration also introduces certain challenges, such as generating false positives. These false positives included synonymous terms like "vegf trap" and "aflibercept," terminologies do not present in the gold timelines such as "aldesleukin," and duplicated abbreviations with different meanings.

Our implemented preprocessing procedure and filtering step effectively reduced false positives not only from the integrated dictionary but also from cases prone to misclassification by the TLINK model. However, this also led to the exclusion of some true pairs from the gold timelines. For example, some patients only had "RAD" files, which do not pertain to chemotherapy details,

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

resulting in missed records that impact our evaluation significantly. Additionally, some chemotherapy pairs were solely mentioned in the plan section, such as "we will give chemo cycle 2 today." While we expected subsequent confirmed notes, they were not present, resulting in the omission of such pairs from our analysis.

Reviewing the test results (Table 3), we obtained the most favorable outcomes for breast cancer, which is the group with the largest sample size. Conversely, the small size of the ovarian cancer type test set poses challenges, as even slight variations in missed or additional pairs can lead to substantial variance. Furthermore, we observed that the gold timeline may not always be entirely accurate, potentially resulting in the omission of rare chemotherapy terms. Addressing these challenges necessitates a larger and more diverse patient dataset in future evaluations.

4 Conclusion and future work

This paper details our efforts in the Chemotimelines 2024 subtask2, focusing on the development of an end-to-end system for chemotherapy timeline extraction. Our experiments utilizing clinical notes from breast, ovarian cancer, and melanoma cases have demonstrated the enhancements made to our pipeline. These enhancements include expanding the system's capabilities by leveraging the UMLS database and implementing preprocessing and directional filtering procedures to effectively reduce false positives.

Future works could potentially include firstly creating a more detailed and precise dictionary using the UMLS, with specific terms tailored to different cancer types, and establishing a synonymous dictionary to prevent duplication of terms. Secondly, it is crucial to exercise caution when removing files such as "RAD" and "SP," especially in cases where patients only possess these notes. Finally, exploring the use of ChatGPT and appropriate prompts as an alternative to the TLINK classifier, which is currently fine-tuned from PubMedBERT, would be a valuable exercise. ChatGPT's superior understanding of sentence context could prove beneficial for those classifications that do not require specific domain knowledge.

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

Acknowledgments

This project has been made possible in part by grant U01CA247760 and U01CA281902 to KC from National Cancer Institute.

References

Gu, Yu, Tinn, Robert, Cheng, Hao, Lucas, Michael, Usuyama, Naoto, Liu, Xiaodong, Naumann, Tristan, Gao, Jianfeng, & Poon, Hoifung. (2022). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, *3*(1), 1–23. https://doi.org/10.1145/3458754

*Jiarui Yao, *Harry Hochheiser, WonJin Yoon, Eli Goldner, & Guergana Savova. (2024). Overview of the 2024 Shared Task on Chemotherapy Treatment Timeline Extraction. *Proceedings of the 6th Clinical Natural Language Processing Workshop*.

Laparra, Egoitz, Xu, Dongfang, & Bethard, Steven. (2018). From Characters to Time Intervals: New Paradigms for Evaluation and Neural Parsing of Time Normalizations. *Transactions of the Association for Computational Linguistics*, *6*, 343–356. https://doi.org/10.1162/tacl_a_00025

Moharasan, Gandhimathi, & Ho, Tu-Bao. (2019). Extraction of Temporal Information from Clinical Narratives. *Journal of Healthcare Informatics Research*, *3*(2), 220–244. https://doi.org/10.1007/s41666-019-00049-0

Najafabadipour, Marjan, Zanin, Massimiliano, Rodríguez-González, Alejandro, Torrente, Maria, Nuñez García, Beatriz, Cruz Bermudez, Juan Luis, Provencio, Mariano, & Menasalvas, Ernestina. (2020). Reconstructing the patient's natural history from electronic health records. *Artificial Intelligence in Medicine*, *105*, 101860. https://doi.org/10.1016/j.artmed.2020.101860

Savova, Guergana K., Masanz, James J., Ogren, Philip V., Zheng, Jiaping, Sohn, Sunghwan, Kipper-Schuler, Karin C., & Chute, Christopher G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association : JAMIA*, *17*(5), 507–513. https://doi.org/10.1136/jamia.2009.001560

Sun, Weiyi, Rumshisky, Anna, & Uzuner, Ozlem. (2013). Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association : JAMIA*, *20*(5), 806–813. https://doi.org/10.1136/amiajnl-2013-001628

UzZaman, Naushad, Llorens, Hector, Allen, James, Derczynski, Leon, Verhagen, Marc, & Pustejovsky, James. (2014). *TempEval-3: Evaluating Events, Time Expressions, and Temporal Relations* (arXiv:1206.5333). arXiv. https://doi.org/10.48550/arXiv.1206.5333

Wang, Liwei, Wampfler, Jason, Dispenzieri, Angela, Xu, Hua, Yang, Ping, & Liu, Hongfang. (2020). Achievability to Extract Specific Date Information for Cancer Research. *AMIA Annual Symposium Proceedings*, *2019*, 893–902. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7153063/

Wang, Yanshan, Wang, Liwei, Rastegar-Mojarad, Majid, Moon, Sungrim, Shen, Feichen, Afzal, Naveed, Liu, Sijia, Zeng, Yuqun, Mehrabi, Saeed, Sohn, Sunghwan, & Liu, Hongfang. (2018). Clinical Information Extraction Applications: A Literature Review. *Journal of Biomedical Informatics*, *77*, 34–49. https://doi.org/10.1016/j.jbi.2017.11.011

Xu, Dongfang, Laparra, Egoitz, & Bethard, Steven. (2019). Pre-trained Contextualized Character Embeddings Lead to Major Improvements in Time Normalization: A Detailed Analysis. In Rada Mihalcea, Ekaterina Shutova, Lun-Wei Ku, Kilian Evang, & Soujanya Poria (Eds.), *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)* (pp. 68–74). Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-1008