Typed Markers and Context for Clinical Temporal Relation Extraction

Cheng Cheng CCHENG2@CMU.EDU
Carnegie Mellon University
Pittsburgh, PA, United States

Jeremy C. Weiss JEREMY.WEISS@NIH.GOV
National Library of Medicine
Bethesda, MD, United States

Abstract

Reliable extraction of temporal relations from clinical notes is a growing need in many clinical research domains. Our work introduces typed markers to the task of clinical temporal relation extraction. We demonstrate that the addition of medical entity information to clinical text as tags with context sentences then input to a transformer-based architecture can outperform more complex systems requiring feature engineering and temporal reasoning. We propose several strategies of typed marker creation that incorporate entity type information at different granularities, with extensive experiments to test their effectiveness. Our system establishes the best result on I2B2, a clinical benchmark dataset for temporal relation extraction, with a F1 at 83.5% that provides a substantial 3.3% improvement over the previous best system. [1]

1. Introduction

Temporal information is essential in clinical settings for understanding disease progression, for providing diagnosis and treatment suggestions, and for evaluating treatment efficacy (Keravnou, 1991; Zhou and Hripcsak, 2007; Nikfarjam et al., 2013). It can provide insight into the relevance of existing medical conditions to a clinical problem, distinguish acute from chronic illness, and disambiguate among potential pathophysiological mechanisms. As an example of temporal disambiguation, a drug administered before a rash develops raises the possibility of drug rash, while a drug administered after a rash develops could indicate treatment for an underlying condition associated with rash. Abundant temporal information exists in clinical notes, making it a crucial to extract them into structured knowledge, enabling patient timeline construction (Viani et al., 2019), providing better clinical decision support (Augusto, 2005), and facilitating time-oriented medical question answering system (Zhou and Hripcsak, 2007).

This paper focuses on the task of temporal relation extraction (TRE) from clinical notes. The extracted temporal orders of medical events have a wide range of use cases: For disease identification, the temporal orders help a physician distinguish between primary and secondary disorders, e.g. ruling out heparin-induced thrombocytopenia (HIT) from other coagulopathies, and identifying syndromes like irritable bowel syndrome from alternating

1. Code for our system is available at https://github.com/Cheng-Cheng2/BTR-CE

episodes of constipation and diarrhea; For clinical phenotyping, temporal orderes enable doctors to distinguish between primary psychiatric disease and substance abuse disorder. For cohort selection in many clinical studies, temporal orders are used to define the inclusion criteria such as in case series and case-control studies as (i) having the exposure, (ii) having the disease, and (iii) the exposure precedes the disease. Whereas the utilization of a general named entity recognition system from the general NLP community could extract entities like a disease or a symptom in clinical text, for a medical entity extraction system, the failure to accurately extract the temporal relations would hinder the use cases in facilitating a doctor's evaluation of a patient's medical condition and improving the quality of care.

It is a nontrivial effort to reliably extract clinical temporal relations on account of the small size of clinical corpora relative to other text corpora, the paucity of high quality notes, the presence of medical jargon in clinical notes, etc. To tackle this task, the Sixth Informatics for Integrating Biology and the Bedside (I2B2) NLP Challenge for Clinical Records (Sun et al., 2013a) introduced a benchmark dataset for clinical temporal relation extaction. Discharge summaries for 310 patients from Partners Healthcare and the Beth Israel Deaconess Medical Center were annotated, focusing on identifying clinical events (EVENT), time expressions (TIMEX or SECTIME), and their temporal relations relative to one another (TLINK). A parallel effort Clinical TempEval (Bethard et al., 2015, 2016, 2017) provided 600 clinical notes and pathology reports from cancer patients at Mayo Clinic, also with EVENT, TIMEX, and TLINK annotated. Our investigation focuses on TRE in I2B2 (and not on TempEval due to the lack of data availability).

Previous works on TRE developed extractor systems that emphasized the integration of syntactic information and temporal reasoning (Zhang et al., 2021; Han et al., 2020; Mathur et al., 2021). However, the medical EVENT, TIMEX, and SECTIME types were not integrated into the systems for clinical TRE, in part because the general English corpora for TRE, such as TB-Dense (Cassidy et al., 2014), do not have additional token types other than syntactic information such as part of speech (POS). On the contrary, clinical temporal annotation schemas (Savova et al., 2009; Sun et al., 2013a; Bethard et al., 2015) have a consensus of clearly annotating clinical EVENT with types ranging from clinically relevant states, procedures, occurrences, and changes, driven by the medical community's need to build automated extraction systems able to distinguish between different event types, and by the need to construct a patient's medical timeline for downstream clinical tasks.

Our investigation attempts to use these available medical EVENT types to improve performance in TRE. To do so, we draw upon research from the natural language processing (NLP) community in general relation extraction (RE), where recent works have shown that a simple technique of using typed markers created from named entity types can outperform more complex architectures that require syntactic graphs or information integrated via auxiliary loss (Zhong and Chen, 2021; Ye et al., 2022; Zhang et al., 2021) for BERT-based models. Specifically, we draw our idea from (Zhong and Chen, 2021) to create typed markers using EVENT, TIMEX, and SECTIME types and use these marked token sequences with context to predict temporal relations.

To the best of our knowledge, our work is the first to introduce typed markers into the task of TRE. We call our best method BTR-CE, as it uses **B**ERT for **t**emporal **r**elation extraction with **c**ontext sentences and **e**vent markers. BTR-CE establishes the best results

on I2B2 with a F1 of 83.5%, providing a substantial 3.3% improvement over the previous best system (Zhang et al., 2021).

Generalizable Insights about Machine Learning in the Context of Healthcare

Our work provides a method to efficiently make use of the typed information for improving results of the clinical TRE task. Although our network architecture is extremely simple with just a BERT encoder and a FFN layer, as typed markers are injected into the token sequence with context sentences prior to training, it manages to outperform more complex BERT-based system making use of temporal reasoning through auxilary loss, and earlier networks making use of typed information through heavy feature engineering, training separate classifiers, and structured learning.

The idea behind our work, the early injection of clinical entity information through markers to help ML systems perform better relation reasoning, can be extended to other clinical prediction tasks where clinical notes are given as input. One potential new application of our typed markers is in the task of understanding clinical progress note[2], where systems are developed to understand the causal relation between assessment and plan subsections from progress note. We hypothesize that injecting typed markers of medical entities into the progress notes will facilitate the ML systems to reason about whether the plan subsections associate with the diagnosis or problem in the assessment.

2. Related Work

Both the clinical NLP and general NLP communities have conducted research on the TRE problem. Earlier systems utilize feature-engineering based statistical models and train classifiers such as in Cherry et al. (2013); Tang et al. (2013); Nikfarjam et al. (2013). Recurrent neural network based models employ an attention mechanism and the integration of integer programming to incorporate structured reasoning (Liu et al., 2019; Han et al., 2019; Leeuwenberg and Moens, 2017). More recently, BERT-based language models (Devlin et al., 2019) have been setting the new state-of-the-art results for a variety of NLP tasks, including in TRE. CTRL-PG (Zhou et al., 2021) developed a probabilistic soft logic regularizer for soft-tuning BERT-based model, which is the present state-of-the-art model in terms of performance on the I2B2 dataset. SGT (Zhang et al., 2021) utilized a syntax-guided graph to improve the supervised training of BERT. Tan et al. (2021) showed that the introduction of hyperbolic geometry boosts the temporal extraction performance on general NLP datasets. TIMERS (Mathur et al., 2021) built a rhetorical and syntactic aware model for document-level temporal relation extraction.

In the NLP literature for the general RE task, Soares et al. (2019) showed that adding [BLANK] markers surrounding entities to input to BERT-based models could outperform methods where relation representations are learned with supervision from a knowledge graph. Zhong and Chen (2021) introduced the **typed markers** created using named entity information which improved upon the RE performance in Soares et al. (2019). Ye et al. (2022) further developed the packed typed marker that better accounts for interrelations between spans with a neighborhood-oriented packing strategy.

2. https://n2c2.dbmi.hms.harvard.edu/2022-track-3

Our method is most similar to that of Zhong and Chen (2021), where we also create typed markers for our TRE task. Whereas Zhong and Chen (2021) aimed to address the general RE task on English domains, where all relations are between named entities from the same type set, our work focus on the clinical TRE task, where TLINKs (temporal relations) are between EVENTs, TIMEXs, and SECTIMEs, each with a different type set. We use a combination of elements from these three type sets to create our typed markers. Also whereas each pair of subject and object entity spans in Zhong and Chen (2021) comes from the same sentence, our subject and object spans can be from sentences anywhere in a document, and to account for this, we respectively concatenate contiguous contextual sentences when creating the input sequences.

3. Methods

Our system mainly consists of three parts: (i) a mechanism to create typed markers to modify the input sequence (Section 3.1), (ii) a BERT-based temporal relation extraction model (Section 3.2), and (iii) context window creation (Section 3.3).

3.1. Typed Marker Pairs

Like Zhong and Chen (2021), we use entity types to enrich the input token sequence for predicting each relation pair. Let span s represent a entity in our notes. We use TLINK(s_i, s_j)=r to denote that subject span s_i and object span s_j has temporal relation r, *e.g.*, Admission (s_i) is before (r) blood test (s_j). Unlike Zhong and Chen (2021) where all entity types come from a single named entity set, we have three different type sets, the EVENT set, TIMEX set, and SECTIME set (Table 2). EVENT represents a medical expression, TIMEX represents a non-section time expression, and SECTIME represents admission or discharge. We also use TIME to represent a time expression in general, either a TIMEX or a SECTIME. TLINK represents a temporal relation, which can be between EVENT-EVENT, EVENT-TIME, and TIME-TIME.

The TRE task is to predict the temporal relation r given a subject span s_i and an object span s_j. Given an input token sequence $T = [t_1, t_2, ...,]$ where $s_i, s_j \subseteq T$, we create subject/object (S/O) typed markers and insert them around the subject and object spans respectively to get a modified sequence:

$$\tilde{T} = [..., [S{:}type_i], s_i, [/S{:}type_i], ..., [O{:}type_j], s_j, [/O{:}type_j], ...] \quad (1)$$

where $type_i, type_j$ correspond to the types of s_i, s_j respectively (Table 2).

Table 1: Medical EVENT, TIMEX, SECTIME types.

	types
EVENT	problem, test, treatment, occurrence, evidential, clinical department
TIMEX	date, duration, frequency, time
SECTIME	admission, discharge

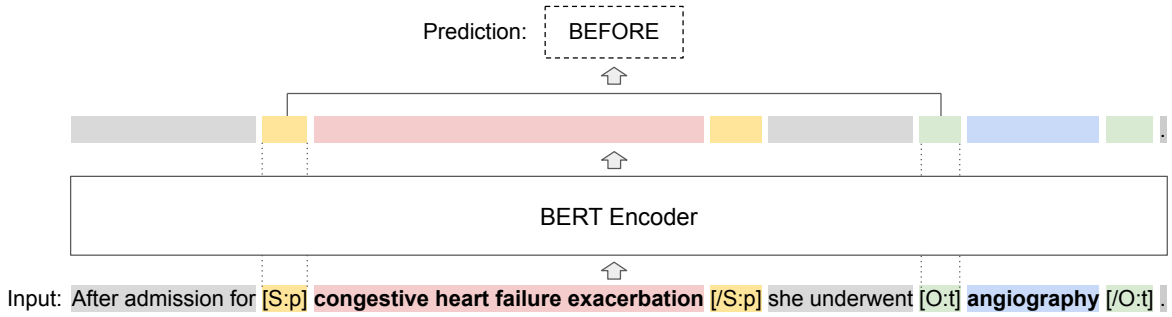

Figure 1: An example from the I2B2 dataset. For the input sentence "After admission for **congestive heart failure exacerbation** she underwent **angiography**.", where events E1="congestive heart failure exacerbation" (a problem), E2="angiography" (a test), and TLINK(E1, E2)=BEFORE, our method creates typed marker pairs [S:p] and [/S:p] (p=problem) for E1, and [O:t] and [/O:t] (t=test) for E2. The encoded representations for [S:p] and [S:t] are used to predict the temporal relation BEFORE.

We demonstrate our workflow with a heart failure example from I2B2 (Figure 1). Given the input sentence "After admission for **congestive heart failure exacerbation** she underwent **angiography**.", where EVENT E1="congestive heart failure exacerbation" (type=problem), EVENT E2="angiography" (type=test), and TLINK(E1, E2)=BEFORE, our method creates typed marker pairs [S:problem] and [/S:problem] for E1, and [O:test] and [/O:test] for E2. The encoded representations for [S:problem] and [S:test] are used to output the final prediction given the ground-truth label BEFORE. The intuition is that allowing the model to discern, for example, that "congestive heart failure exacerbation" is a medical problem and that "angiography" is a medical test informs the model to predict that "congestive heart failure exacerbation" precedes "angiography", since a patient manifesting an acute medical problem will likely be administered a medical test.

3.1.1. Different strategies of Typed Marker Creation

We tried different strategies for creating the typed markers to explore how different granularity of fusion of entity types affect the TRE results:

1. **DEFAULT**: for the simplest default setup we use "EVENT" as the type for EVENT span, "TIMEX" as the type for TIMEX span, and the two types, "admission" and "discharge" (Table 2), for SECTIME span.

2. **EVENT**: we use the EVENT types in Table 2 for EVENT span, and keep the rest of the types the same as in DEFAULT (Strategy 1).

3. **EVENT + POS**: we add part-of-speech (POS) tags to the marker types set in EVENT (Strategy 2), creating a composite type for each span as "EVENT type: POS type".

4. **EVENT + TIMEX**: we use the TIMEX types in Table 2 for TIMEX span and keep the rest of the types the same as in EVENT (Strategy 2).

5. **EVENT + TIMEX + POS**: we add POS tags to the marker types set in EVENT+TIMEX (Strategy 4), creating a composite type for each span as "EVENT+TIMEX type: POS type".

Table 2: Different strategies of Typed Marker Creation.

strategy	types
DEFAULT	"EVENT"", "TIMEX", and SECTIME types
EVENT	EVENT types + DEFAULT
EVENT+POS	part-of-speech (POS) tags + EVENT
EVENT+TIMEX	TIMEX types + EVENT
EVENT+TIMEX+POS	"EVENT+TIMEX type: POS type"

3.2. Relation Extraction Model

Our relation extraction model is straightforward and similar to Zhong and Chen (2021). Let I_i and I_j represent the index of [S:$type_i$] and [O:$type_j$] in $\tilde{T}$ in Equation 1. The modified sequence $\tilde{T}$ is first encoded by BERT as:

$$H = \text{BERT}(\tilde{T}). \tag{2}$$

The concatenation of the encoded markers for [S:$type_i$] and [O:$type_j$] is represented as $M = [h_{I_1}; h_{I_2}]$. M is passed through a linear layer to predict the temporal relation as $\tilde{r}$:

$$\hat{r} = \text{Softmax}(\text{Linear}(\text{M})). \tag{3}$$

.

The cross entropy loss is used to train the network as:

$$\mathcal{L}(\hat{r}, r) = \sum_{k \in K} r^k \log \hat{r}^k \tag{4}$$

where K is the label set for temporal relation r.

3.3. Context Window

Previous works (Wadden et al., 2019; Zhang et al., 2021) have shown that adding context sentences may boost performance for entity and relation extraction tasks. We follow Wadden et al. (2019) to add a W-sentence context window to the left side of the left span and to the right side of the right span, and also add a V-sentence context window to the right side of the left span, and to the left side of the right span. In definition, for subject $s_i \subseteq N_{O(s_i)}$ and object $s_j \subseteq N_{O(s_j)}$ where $O(s_x)$ represents the index of the sentence $N_{O(s_x)}$ in the document where s_i and s_j are located, we create the input sequence T as:

$$\begin{aligned} T = \{&N_{\min(O(s_i),O(s_j))-W} \ldots, N_{\min(O(s_i),O(s_j))}, \ldots, N_{\min(O(s_i),O(s_j))+V}, \\ &N_{\max(O(s_i),O(s_j))-V}, \ldots, N_{\max(O(s_i),O(s_j))} \ldots, N_{\max(O(s_i),O(s_j))+W}\} \end{aligned} \tag{5}$$

V and W are treated as training hyperparameters.

4. Cohort

Data and Code Availability This paper uses two benchmark datasets (Table 3): 1. I2B2 which is available on the National NLP Clinical Challenges (Sun et al., 2013a); it has three temporal relation classes, AFTER (9.7%), BEFORE (53.3%), and OVERLAP (37.0%). 2. TB-Dense which is made available by Cassidy et al. (2014) based on annotations on the TimeBank Corpus (Pustejovsky et al., 2003); It has six temporal relation classes, AFTER (18.4%), BEFORE (22.1%), SIMULTANEOUS (1.5%), INCLUDES (4.5%), IS_INCLUDE (5.7%), and VAGUE (47.7%). I2B2 contains clinical events and temporal relations annotated on clinical discharge summaries, provided as 190 training documents and 120 test documents, and we follow Zhang et al. (2021) to use 9 of the training documents as a development set. TB-dense is a general NLP dataset for temporal relation extraction, the train/development/test sets are given. Although our main analysis focuses on the clinical TRE task with I2B2, we use TB-Dense in attempts to demonstrate that even though a general English corpus, unlike the clinical corpus, does not contain rich entity type information, typed markers using POS alone can still effectively outperforms a plain BERT.

Table 3: Cohort table.

Dataset		Train	Dev	Test
I2B2	doc	181	9	120
	relation	32,508	1,259	27,735
TB-Dense	doc	22	5	9
	relation	4,032	629	1,427

4.1. Data Preprocessing

To address the data imbalance problem at training time, we augment the minority class in data by flipping the order of two events and add a TLINK to the data. For I2B2 we flip links with BEFORE as relation to augment AFTER (the minority class of 9.7% out of total), and for TB-Dense we flip SIMULTANEOUS (the minority class of 1.5% out of total) to augment itself.

5. Results

5.1. Evaluation Approach/Study Design

5.1.1. Experimental design

For our BERT-base encoder we follow the exact architecture as Devlin et al. (2019), with a dropout rate as 0.1 and max sequence length as 512. To train our model we also utilize an Adam optimizer with a warmup ratio as 0.1 (Loshchilov and Hutter, 2017; Kingma and Ba, 2014) and set training epochs as 20. We perform grid search on a development set for other hyper-parameters including learning rate λ from $\{10^{-5}, 2 \times 10^{-5}, 10^{-4}\}$, context window size W from $\{0, 1, 2, 3\}$, and train batch size B from $\{8, 16, 32\}$. Empirically we identify the best development set performance, with $\lambda = 2 \times 10^{-5}$ and $B = 16$ for both

datasets, and $W = 2$ for I2B2 and $W = 0$ for TB-Dense. Also we fix $V = 2$ for I2B2 because concatenating everything in-between causes the sequence to exceed the max length of BERT input (512) and this avoids truncation. And $V = 0$ for TB-Dense because all span pairs come from the same given text sequence. For each of our experiments with typed markers, we run five experiments with different random seeds and aggregate the results.

5.1.2. Evaluation Metrics

For fair comparison, we follow the exact evaluation strategy of Zhou et al. (2021) for both I2B2 and TB-dense. For both datasets, all values reported using our method are aggregated from five experiments with different random seeds.

I2B2 Average precision, recall, and F1 are computed using the evaluation script provided by Sun et al. (2013a), where *closure* (UzZaman and Allen, 2011) is employed to account for transitivity of relations, *e.g.*, if there exists a TLINK(E1, E2)=BEFORE, and TLINK(E2, E3)=BEFORE, than closure adds a derived TLINK(E1, E3)=BEFORE. Different from standard machine learning precision and recall metrics, here precision for one patient is calculated as the percentage of system predicted TLINKs that can be verified in the closure graph of ground truth TLINKs, and recall is calculated as the percentage of ground truth TLINKs that can be verified in the closure graph of the system predicted TLINKs. The F1 score is the harmonic mean of precision and recall.

TB-dense We compute the standard micro average precision, recall, and F1 for only the Event-Event TLINKs following Zhou et al. (2021); Han et al. (2020); Meng and Rumshisky (2018). Note that under multiclass classification the micro average precision, recall, and F1 are equal (Grandini et al., 2020).

5.1.3. Baseline Methods

We compare against both feature-engineering based methods from systems submitted to the I2B2 2012 challenge and more recent neural-network methods.

I2B2 (1) Feature-engineering based statistic models. MaxEnt-SVM (Cherry et al., 2013) utilizes a mixture of Max Entropy Classifier and Support Vector Machine (SVM). CRF-SVM (Tang et al., 2013) utilizes a hybrid system that trains different classifiers for different link types based on Conditional Random Fields and SVM. RULE-SVM (Nikfarjam et al., 2013) adds a rule based component to their SVM based system. (2) Neural network based model. RNN-ATT (Liu et al., 2019) utilizes a recurrent neural network (RNN) with an attention layer. SP-ILP (Han et al., 2019; Leeuwenberg and Moens, 2017) employs a recurrent neural network jointly trained with structured learning and integer programming constraints. HyperGeo (Tan et al., 2021) applies a RoBERTa-base to embed events into hyperbolic space. CTRL-PG (Zhou et al., 2021) utilizes a BERT with probabilistic soft logic regularization and global inference. BTR-C uses a BERT temporal relation extractor with context window. For fair comparison, we collect results for RULE-SVM, MaxEnt-SVM, and CFR-SVM from Cherry et al. (2013); Tang et al. (2013); Nikfarjam et al. (2013), for RNN-ATT, SP-ILP, and CTRL-PG from Zhou et al. (2021), and run experiments for HyperGeo (without external knowledge) and BTR-C with the same network setup in Section 5.1.1.

TB-Dense CAEVO (Leeuwenberg and Moens, 2017) proposes a cascading event ordering architecture for rule-based classifiers. LSTM-DP (Cheng and Miyao, 2017) adopts a bidirectional long short-term memory (Bi-LSTM) for incorporating the dependency tree. GCL (Meng and Rumshisky, 2018) adds a global context layer to a RNN-based structure. SGT (Zhang et al., 2021) utilizes a syntax-guided attention mechanism to incorporate the dependency parsing tree for improving the supervised training of BERT. SP-ILP, BTR-C, and CTRL-PG are also compared against for I2B2. We collect the results for CAEVO, LSTM-DP, GCL, SP-ILP from Han et al. (2019), for CTRL-PG from Zhou et al. (2021), and for SGT from Zhang et al. (2021), and run experiments for BTR-C.

Table 4: Experimental results across methods on I2B2 test set (%).

	P	R	F1
RULE-SVM	71.1	58.4	64.1
MaxEnt-SVM	75.4	64.6	69.5
CFR-SVM	72.3	66.8	70.1
RNN-ATT	71.9	69.1	70.5
SP-ILP	78.1	78.2	78.2
HyperGeo	79.3	80.1	79.7
CTRL-PG	**86.8**	74.5	80.2
BTR-C	80.8	82.0	81.3
BTR-CE	83.0	**83.9**	**83.5**

Table 5: Ablation study of our best strategy.

	P		R		F1		F1 diff
	mean	std	mean	std	mean	std	
BTR-CE	**83.0**	0.6	**83.9**	0.4	**83.5**	0.5	–
w/o EVENT	80.8	0.6	82.0	0.2	81.3	0.4	2.2
w/o CW	61.6	0.2	66.0	0.4	63.7	0.3	19.8
w/o CW or EVENT	61.5	0.8	65.5	0.9	63.4	0.8	20.1

5.2. Results on I2B2

We experiment with different strategies for typed marker creation (Section 3.1.1) and the EVENT strategy (using EVENT types, SECTIME types, and "TIMEX" for TIMEX span) results in the best performance across precision (83.0%), recall (83.9%), and F1 (83.5%) (Table 6). Our best method BTR-CE, as it uses **B**ERT for **t**emporal **r**elation extraction with **c**ontext sentences and **e**vent makers. Comparing our BTR-CE against all the baseline methods (Table 4), it achieves the highest recall and F1, and the second highest precision across methods, with a remarkable 3.3% F1 improvement compared to CTRL-PG, the previous state-of-the-art model on I2B2.

Table 6: Experimental results across different strategies of typed marker creation on I2B2 test set (%).

Strategy	P		R		F1	
	mean	std	mean	std	mean	std
DEFAULT	82.9	0.6	83.7	0.4	83.3	0.5
EVENT+POS	82.7	0.3	83.6	0.2	83.2	0.3
EVENT+TIME	82.8	0.4	83.6	0.4	83.2	0.4
EVENT+TIME+POS	82.9	0.3	83.7	0.3	83.3	0.3
EVENT	**83.0**	0.6	**83.9**	0.4	**83.5**	0.5

Performance Breakdown for BTR-CE. We perform subgroup analysis based on temporal relation types (Table 7), and link types (Table 8) for our best method. Performance breakdown based on temporal relation types (Table 7) demonstrates that the AFTER type has the lowest score across metrics (67.3% as precision, 65.4% as recall, and 66.3% as F1), while the BEFORE type achieves the best precision and F1 (88.7% and 86.1%). Their ascending order of overall TRE performance measured by F1 is consistent with their increasing proportion of the data, ranging from the AFTER type making up 9.7% of the I2B2 test data to the BEFORE type making up 53.2%. The performance breakdown based on link types reveals a similar trend of performance increment with higher data proportion. The TIMEX-TIMEX type, making up only 0.8% of the data, has the worst performance across metrics (73.3% as precision, 73.6% as recall, and 73.4% as F1), while the Event-Timex type (with SECTIME), making up 48.0% of the data, has the best performance across metrics (82.6% as precision, 82.7% as recall, and 82.6% as F1).

Ablation for BTR-CE. Per ablation study (Table 5), we observe that removing EVENT markers leads to a 2.2% drop in F1, removing context window (CW) leads to a striking 19.8% drop in F1, and removing both (with a BERT extractor alone) leads to a 20.1% drop.

Table 7: Performance breakdown based on temporal relation types on I2B2 test set for BTR-CE. P, R, and F1 are measured in percentage (%).

Temporal Relation type	Count (%)	P	R	F1
ALL	27,735 (100.0)	83.0	83.9	83.5
AFTER	2,729 (9.8)	67.3	65.4	66.3
OVERLAP	9,893 (35.7)	82.9	88.6	85.7
BEFORE	15,113 (54.5)	88.7	83.6	86.1

Table 8: Performance breakdown based on link types on I2B2 test set for BTR-CE. E stands for EVENT, T stands for TIME (either a TIMEX or a SECTIME), and SEC stands for SECTIME. P, R, and F1 are measured in percentage (%).

Link Type	Count(%)	P	R	F1
ALL	27,735 (100.0)	83.0	83.9	83.5
T-T	214 (0.8)	73.3	73.6	73.4
E-T - no SEC	2,452 (8.8)	82.6	82.8	82.7
E-T - with SEC	13,317 (48.0)	84.7	88.5	86.5
E-E	11,752 (42.4)	82.6	82.7	82.6

5.3. Results on TB-Dense

Unlike the clinical TRE task where benchmark datasets have EVENT, TIMEX type information (Sun et al., 2013a; Bethard et al., 2015, 2016, 2017), TB-Dense, the benchmark dataset for the general English TRE task, does not have rich entity type information besides POS tags, so we only could create typed markers for TB-Dense using POS tags, we call it BTR-CP. While the F1 (62.9%) from BTR-CP improves upon the F1 (61.5%) from BTR-C, a plain BERT based model without typed markers by 1.4%, it cannot outperform SP-ILP, CTRL-PG and SGT, the complex neural network based methods where temporal reasoning and dependency parsing tree are incorporated (Appendix Table A1).

6. Discussion

Our BTR-CE, using both context window and EVENT markers, establishes the new state-of-the-art performance for the I2B2 clinical temporal extraction task, with a 3.3% performance improvement upon the previous best model CTRL-PG (Table 4). Compared to previous neural network based systems, our method is simpler yet effective since both typed markers and context-window only modify the input, unlike SP-ILP that requires structured learning and integer programming, or CTRL-PG that needs an auxiliary loss, or HyperGeo that needs the non-euclidean projection of the network layers on top of a transformer. BTR-CE uses a simple FFN layer on top of BERT to achieve the current best performance, with typed markers and context window constructed prior to training. Even our simplest marker creation strategy, the DEFAULT strategy (Section 3.1.1) where we only use "EVENT" for EVENT span, "TIMEX" for TIMEX span, and "admission" and "discharge" for SECTIME span, achieves a high F1 score of 83.3%, outperforming CTRL-PG (Table 6). Whereas other systems also explore the idea of injection of type information, as SVM-CRF learns different classifiers for different link types and SP-ILP uses structured learning to separate document-level vs entity-level relations, our injection of the type information at the input layer, even the DEFAULT form that only separates EVENT, TIMEX, and SECTIME, is both simple and high-performing. Notably the use of context window is very important for I2B2 (Table 5), but not for TB-Dense (grid search on development set chooses W=0), highlighting the importance of context in clinical TRE. In our released code we include detail for context

window construction since we find such information is kind of overlooked in both manuscript and code repository from previous research, making it hard for reproduciblity.

From the performance breakdown by relations types for I2B2 (Table 7), we observe that TLINKs with an AFTER relation perform relatively poorly with a F1 of 66.3% (Table 7), while both BEFORE and OVERLAP relations perform relatively well with F1 scores more than 85%. This could be due to a class imbalance problem, as AFTER links make up only 9.7% of the total I2B2 data. Although we attempt to augment the AFTER links in train data by flipping the BEFORE links, the final results still demonstrate need for improvement. This trend that scarce data types have worse performance can also be observed in the performance breakdown based on link types (Table 8) , as TIME-TIME links, making up only 0.8% of the data, has the worst performance at 73.4% for F1, while all other links have F1 more than 82.5%. Note this trend is opposite to that of CTRL-PG, as they build their global inference (GI) algorithm based on their stated belief that the TIME-TIME links are the easiest to predict, and their GI unit is shown to improve their F1 performance by 1.8%.

Our typed marker results on TB-Dense (Appendix Table A1) are not as strong as on I2B2 and do not surpass the performance of previous studies. We posit that this is due to having only the POS entity type available in TB-Dense, unlike I2B2 where we incorporate medical typing in EVENT and SECTIME markers alongside TIMEX and POS. Reviewing our results, we see that the inclusion of marker type (DEFAULT strategy) and specifically EVENT markers provides substantial benefit for the clinical TRE task, and these can be extracted using off-the-shelf systems to construct the typed markers in our pipeline. On TB-Dense, although typed markers do improve results (F1 of 61.5%) compared to BTR-C (F1 62.9%), the BERT Base model, their addition does not outperform other BERT-based methods like CTRL-PG (F1 of 65.2%, Zhou et al. (2021)), where temporal reasoning is incorporated through an auxiliary loss, or SGT (F1 of 67.1%, Zhang et al. (2021)) where a syntax-guided attention mechanism incorporates the dependency parsing tree.

In future work, it is worth exploring if the incorporation of temporal reasoning and dependency tree into our system can further improve performance. We will also consider generalizing and evaluating our method on out-of-domain datasets, *e.g.*, if provided access to TempEval (Bethard et al., 2015, 2016, 2017).

Limitations Our method under-performs for minority classes when class imbalance is present, and future study of temporal relations should work towards improving performance of the minority classes (the AFTER type) and links where data are scarce (the TIME-TIME type). Also, while our system tackles the clinical TRE task, where EVENT and TIME spans and types are given, in the future one could consider constructing end-to-end entity and relation extraction system like Zhong and Chen (2021) for clinical TRE and measure the performance degradation when entity spans and types are predicted rather than given.

Our study also possesses limitations in temporal representation. In TRE the retrieved event temporal orders are relative, commonly losing partial information of the absolute timeline, *e.g.*, events time-located by timestamped intervals. One could incorporate structured data to better recover the absolute patient timeline (like Leeuwenberg and Moens (2020)) in a multimodal framework by combining temporal information from both tabular EHR and the textual temporal relations. Additionally, the original I2B2 annotations used eight temporal relations: BEFORE, AFTER, SIMULTANOUS, BEGUN_BY, ENDED_BY, DUR-

ING, and BEFORE_OVERLAP (Sun et al., 2013b). However, due to low inter-annotator agreement on some relations, BEFORE, ENDED_BY, and BEFORE_OEVERLAP were merged as BEFORE, BEGUN_BY, and AFTER were merged as AFTER, and SIMULTANEOUSE, OVERLAP, and DURING were merged as OVERLAP, resulting in the three label classes for I2B2 (Sun et al., 2013a). This merging process is another source of temporal granularity loss, demonstrating potential merit for reconstructing a more holistic timeline representation for performing effective temporal reasoning.

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Library of Medicine (NLM), National Institutes of Health (NIH). This work utilized the computational resources of the NIH HPC Biowulf cluster.

References

Juan Carlos Augusto. Temporal reasoning for decision support in medicine. *Artificial intelligence in medicine*, 33(1):1–24, 2005.

Steven Bethard, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen. SemEval-2015 task 6: Clinical TempEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 806–814, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.18653/v1/S15-2136.

Steven Bethard, Marine Carpuat, Daniel Cer, David Jurgens, Preslav Nakov, and Torsten Zesch, editors. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, California, June 2016. Association for Computational Linguistics.

Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. SemEval-2017 task 12: Clinical TempEval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2093.

Taylor Cassidy, Bill McDowell, Nathanel Chambers, and Steven Bethard. An annotation framework for dense event ordering. Technical report, CMU Pittsburgh PA, 2014.

Fei Cheng and Yusuke Miyao. Classifying temporal relations by bidirectional LSTM over dependency paths. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–6, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2001.

Colin Cherry, Xiaodan Zhu, Joel Martin, and Berry de Bruijn. A la recherche du temps perdu: extracting temporal relations from medical text in the 2012 i2b2 nlp challenge. *Journal of the American Medical Informatics Association*, 20(5):843–848, 2013.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, Volume 1, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*, 2020.

Rujun Han, I-Hung Hsu, Mu Yang, Aram Galstyan, Ralph Weischedel, and Nanyun Peng. Deep structured neural network for event temporal relation extraction. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 666–106, Hong Kong, China, November 2019. Association for Computational Linguistics.

Rujun Han, Yichao Zhou, and Nanyun Peng. Domain knowledge empowered structured neural net for end-to-end event temporal relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Elpida Keravnou. Medical temporal reasoning. 1991.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Artuur Leeuwenberg and Marie-Francine Moens. Structured learning for temporal relation extraction from clinical records. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1150–1158, Valencia, Spain, April 2017. Association for Computational Linguistics.

Artuur Leeuwenberg and Marie-Francine Moens. Towards extracting absolute event timelines from english clinical reports. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2710–2719, 2020.

Sijia Liu, Liwei Wang, Vipin Chaudhary, and Hongfang Liu. Attention neural model for temporal relation extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 134–139, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1917.

Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2017.

Puneet Mathur, Rajiv Jain, Franck Dernoncourt, Vlad Morariu, Quan Hung Tran, and Dinesh Manocha. Timers: document-level temporal relation extraction. In *Proceedings of the 59th Annual Meeting of the ACL and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 524–533, 2021.

Yuanliang Meng and Anna Rumshisky. Context-aware neural model for temporal information extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 527–536, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1049.

Azadeh Nikfarjam, Ehsan Emadzadeh, and Graciela Gonzalez. Towards generating a patient's timeline: extracting temporal relationships from clinical notes. *Journal of biomedical informatics*, 46:S40–S47, 2013.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK, 2003.

Guergana Savova, Steven Bethard, Will Styler, James Martin, Martha Palmer, James Masanz, and Wayne Ward. Towards temporal relation discovery from the clinical narrative. In *AMIA annual symposium proceedings*, volume 2009, page 568, 2009.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*, 2019.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association*, 20 (5):806–813, 04 2013a. ISSN 1067-5027. doi: 10.1136/amiajnl-2013-001628.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. Annotating temporal information in clinical narratives. *Journal of biomedical informatics*, 46:S5–S12, 2013b.

Xingwei Tan, Gabriele Pergola, and Yulan He. Extracting event temporal relations via hyperbolic geometry. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8065–8077, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

Buzhou Tang, Yonghui Wu, Min Jiang, Yukun Chen, Joshua C Denny, and Hua Xu. A hybrid system for temporal information extraction from clinical text. *Journal of the American Medical Informatics Association*, 20(5):828–835, 2013.

Naushad UzZaman and James Allen. Temporal evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 351–356. Association for Computational Linguistics, June 2011.

Natalia Viani, Hegler Tissot, Ariane Bernardino, and Sumithra Velupillai. Annotating temporal information in clinical notes for timeline reconstruction: Towards the definition of calendar expressions. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 201–210, Florence, Italy, August 2019. Association for Computational Linguistics.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. Entity, relation, and event extraction with contextualized span representations. *arXiv preprint arXiv:1909.03546*, 2019.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4904–4917, Dublin, Ireland, May 2022. Association for Computational Linguistics.

Shuaicheng Zhang, Lifu Huang, and Qiang Ning. Extracting temporal event relation with syntax-guided graph transformer. *arXiv preprint arXiv:2104.09570*, 2021.

Zexuan Zhong and Danqi Chen. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

Li Zhou and George Hripcsak. Temporal reasoning with medical data—a review with emphasis on medical natural language processing. *Journal of biomedical informatics*, 40 (2):183–202, 2007.

Yichao Zhou, Yu Yan, Rujun Han, J Harry Caufield, Kai-Wei Chang, Yizhou Sun, Peipei Ping, and Wei Wang. 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, 2021.

Appendix A. Results on TB-Dense

While the F1 (62.9%) from our method improves upon the F1 (61.5%) from BTR-C, a plain BERT model without typed markers by 1.4%, it cannot outperform SP-ILP, CTRL-PG and SGT, complex neural network based methods where temporal reasoning and dependency parsing tree are incorporated (Appendix Table A1).

Table A1: Experimental results across methods on TB-Dense test set (%).

	F1(=P=R)
CAEVO	49.4
LSTM-DP	52.9
GCL	57.0
SP-ILP	63.2
CTRL-PG	65.2
SGT	**67.1**
BTR-C	61.5
BTR-CP	62.9