Defining and Learning Refined Temporal Relations in the Clinical Narrative

Chen Lin[1] (co-first author), Kristin Wright-Bettner[2] (co-first author),
Timothy Miller[1], Steven Bethard[3], Dmitriy Dligach[4],
Martha Palmer[2], James H. Martin[2] and Guergana Savova[1]

[1]Boston Children's Hospital and Harvard Medical School, Boston, MA
[2]Departments of Linguistics and Computer Science, University of Colorado, Boulder, CO
[3]School of Information, University of Arizona, Tucson, AZ
[4]Department of Computer Science, Loyola University Chicago, Chicago, IL
[1]{first.last}@childrens.harvard.edu
[2]{first.last}@colorado.edu
[3]bethard@arizona.edu
[4]ddligach@luc.edu

Abstract

We present refinements over existing temporal relation annotations in the Electronic Medical Record clinical narrative. We refined the THYME corpus annotations to more faithfully represent nuanced temporality and nuanced temporal-coreferential relations. The main contributions are in re-defining CONTAINS and OVERLAP relations into CONTAINS, CONTAINS-SUBEVENT, OVERLAP and NOTED-ON. We demonstrate that these refinements lead to substantial gains in learnability for state-of-the-art transformer models as compared to previously reported results on the original THYME corpus. We thus establish a baseline for the automatic extraction of these refined temporal relations. Although our study is done on clinical narrative, we believe it addresses far-reaching challenges that are corpus- and domain- agnostic.

1 Introduction

Temporal relation extraction and reasoning in the clinical domain continues to be a primary area of interest due to the potential impact on disease understanding and, ultimately, patient care. A significant body of text available for this purpose is the THYME (**T**emporal **H**istories of **Y**our **M**edical **E**vents) corpus (Styler IV et al., 2014), consisting of 594 clinical and pathology notes on colon cancer patients and 600 radiology, oncology and clinical notes on brain cancer patients, all from the Electronic Medical Record (EMR) of a leading US medical center. This dataset has previously undergone a variety of annotation efforts, most notably temporal annotation (Styler IV et al., 2014). It has been part of several SemEval shared tasks such as Clinical TempEval (Bethard et al., 2017) where state-of-the-art results have been established. Our goal was to utilize this THYME corpus to enable the extraction of more extensive patient timelines by manually creating cross-document links that built off the pre-existing single file annotations.

(Wright-Bettner et al., 2019) discuss that a subset of the THYME temporal annotations contributed to incompatible temporal inferences, thus reducing their ability to support meaningful temporal reasoning. Accuracy and informativeness of temporal relation gold annotations are essential for their effectiveness in training a system for temporal relation extraction. We build on this work by offering an in-depth discussion of three key temporal relations – CONTAINS, CONTAINS-SUBEVENT (abbreviated as CON-SUB), and NOTED-ON – and how the addition of the last two types enhances the learnability of the temporal relations by resolving the conflicting temporal information in the original annotations.

While these revisions are corpus-specific, the reasoning behind them has far-reaching implications for automated timeline extraction. Since the cross-document linking task inherently deals with multiple, discrete narratives, it exposes the practical impact of discourse context on word

Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, pages 104–114
November 20, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

sense (different narratives have different goals, which in turn influences meaning interpretation). This is discussed in detail in Section 4. We empirically found it essential to take changes in discourse context into account and suggest the same would be true for any annotation project that is interested in temporal reasoning, particularly those dealing with longer timelines (i.e., beyond the single-document level).

Recent developments in natural language processing establish neural approaches and more specifically transformer-based methods as the state of the art. Pre-trained models such as BERT (Devlin et al., 2019), BioBERT (Lee et al., 2020), Xlnet (Yang et al., 2019), ALBERT (Lan et al., 2020), RoBERTa (Liu et al., 2019), BART (Lewis et al., 2019), and SpanBERT (Joshi et al., 2020) report significant gains on multiple tasks. Thus, we demonstrate the learnability of the refined temporal relations in the context of these recent methodological developments.

2 Dataset

The 594 notes that make up the colon cancer part of the THYME corpus are grouped into sets, each set pertaining to a single patient and consisting of three notes written at different times during the patient's course of care. These notes had been previously annotated for five different intra-document temporal relations (BEFORE, OVERLAP, BEGINS-ON, ENDS-ON and CONTAINS), a subset of the ISO-TimeML temporal link (TLINK) types (Pustejovsky et al., 2010, Styler IV et al., 2014)[1]. To keep annotation manageable and circumvent massively inferential temporal linking, the THYME guidelines constrained TLINK creation to events within the same sentence or adjacent sentences, and specifically prohibited TLINKing across sections (these are clinically-delineated sections separated from each other by numerical section IDs – *History of Present Illness* is section 20103, *Vital Signs* is section 20110, etc.) Linguistic evidence for creating these TLINKs included local cues such as temporal prepositions and adjectives (e.g., *during*, *subsequent to*, *prior to*), chronological narrative progression, and so forth.

Additionally, the notes had been separately annotated for intra-document coreference (IDENTICAL) and bridging (SET-SUBSET, WHOLE-PART) relations, which were later merged with the temporal annotations (Wright-Bettner et al., 2019). Temporal relations alone are insufficient for timeline extraction; coreference relations are also necessary (is the tumor seen in September the same as the one seen in March?).

Pursuant to our goal of reasoning over large-scale timelines, we built off these pre-existing within-note annotations by manually adding coreference and bridging links across each set of three notes. In the process, we discovered a subset of the original CONTAINS relations contributed to temporally-conflicting information, which led to the addition of two new TLINKs: NOTED-ON and CON-SUB (Wright-Bettner et al., 2019). We discuss below how these updates contribute to more accurate and comprehensive temporal relations which facilitated cross-document linking. As such, this is one of the few studies in clinical NLP for cross-document temporal relation annotations (see Raghavan et al., 2014 and Wright-Bettner et al., 2019; also see Song et al., 2018 for general domain cross-document temporal annotation discussions).

3 Refined Temporal Relation: NOTED-ON

The THYME guidelines specified that tests should always be in a CONTAINS relation with their results, as in (1), observing that it is inferential to say that something seen on the test exists before or after the test.

1. *January 17, 2009:* ***CT*** *abdomen and pelvis shows liver* ***metastases****.*
   a. *CT* CONTAINS *metastases*

We agree with the authors in part, particularly for a project that did not have coreference relations and that relied heavily on explicitly-temporal cues (*after*, *during*, *before*, etc.).

However, once the coreference links had been merged with the temporal annotations, more information was revealed about the temporal nature of the events. For example, *metastases* in

[1] This approach to temporal annotation in fact evolved directly from the ISO-TimeML model, in collaboration with Prof. James Pustejovsky. The Basic Formal Ontology has introduced similar relations (Smith et al., 2005), and a post-hoc mapping could be done between the relations proposed here and those in BFO 2.0. However, we defer to more philosophical scholars to determine the similarities and differences between ISO TimeML and BFO 2.0.

(1) might well be IDENTICAL to a later mention, such as:

2. *February 20, 2009, liver* ***metastases*** *resected with clear margins.*

The merged THYME and coreference annotations[2] for (1) and (2) were as follows:

- *January 17, 2009* CONTAINS *CT*
- *CT* CONTAINS *metastases*
- *February 20, 2009* CONTAINS *resected*
- *metastases* OVERLAPS *resected*
- *metastases* IDENTICAL *metastases*

Together, these relations entail that the same liver abnormalities were temporally contained by January 17, 2009 and temporally overlapped February 20, 2009 – a logical impossibility. This situation was extremely frequent in the data, reducing the informativeness of the CONTAINS links and the timeline as a whole. We therefore manually converted these links to a subtype of OVERLAP relation, NOTED-ON. NOTED-ON conveys temporal overlap between events and additionally says that Event A (*result*) was observed on/by Event B (*test*). (1a) was therefore re-annotated as follows:

3. *metastases* NOTED-ON *CT*

The application of this link was so constrained that it was fairly easy to annotate; we therefore added it as a single-annotated post-processing step.

4 Refined Temporal Relation: CONTAINS-SUBEVENT

4.1 Motivation and annotation

Early in the cross-document annotation pilot, we found the pre-existing intra-document schema categories were insufficient for dealing with cross-narrative phenomena, which in turn led to the addition of the new CON-SUB relation. Consider:

4. **Note A**: 2009: *Patient presents with recurrent* ***adenocarcinoma****. CT abdomen and pelvis ordered for further staging of this colon* ***cancer****.*

i. *adenocarcinoma* IDENT *cancer*

Note B: *Successful removal of primary* ***tumor*** *in 2004. Recurrent* ***adenocarcinoma*** *is inoperable. Patient here today to discuss other treatment options for his colon* ***cancer****.*

ii. *cancer* CONTAINS *tumor*

iii. *cancer* CONTAINS *adenocarcinoma*

Both sets of intra-document links are pragmatically appropriate. Discourse contexts can expand or reduce the level of granularity at which a sense is interpreted (Recasens et al., 2011; also see Hobbs, 1985). In Note A, the text supports a coarse-grained interpretation of *adenocarcinoma*, or what Hovy et al., 2013 term a "wide" reading; it refers generally to the patient's cancer. Note B, however, requires a fine-grained ("narrow") interpretation – *adenocarcinoma* here refers specifically to the new, inoperable tumor and is contrasted with the original, resected tumor.

The quandary for the cross-document task lies in whether to link *adenocarcinoma* in A as IDENTICAL to *adenocarcinoma* in B. An IDENTICAL relation entails logical impossibilities: assuming we also link *cancer*$_{A}$ as IDENT to *cancer*$_{B}$, the combined within- and cross-document relations now say the recurrent adenocarcinoma temporally contains itself *and* the primary tumor which was removed years earlier. This reduces the meaningfulness of the CONTAINS links and therefore the timeline. On the other hand, leaving the two *adenocarcinoma* references unlinked fails to capture the significant semantic relation between two identical strings (*recurrent adenocarcinoma*) that do in fact refer, on some level of granularity, to the same real-world event. In either case, annotators are stymied.

Clearly, the pre-existing intra-document schema categories, specifically the binary coreference choice ($A = B$ or $A \neq B$), were insufficient for dealing with the sense variation and nuanced event structure exposed by multiple narratives. We therefore introduced CON-SUB (based on O'Gorman et al., 2016) as a subtype of the CONTAINS TLINK type. While CONTAINS conveys only temporal containment, CON-SUB *additionally* says that Event B is intrinsically part

[2] Note: This and all examples in this paper have been fabricated for patient privacy. However, they are all linguistically similar to actual sentences from the corpus.

of the structure of Event A[3]. The difference may be seen in (5):

5. *During patient's neoadjuvant* ***treatment****, she was in a car* ***accident*** *which* ***delayed cycle*** *4*.
a. *treatment* CONTAINS *accident*, *delayed*
b. *treatment* CON-SUB *cycle*

We were then able to re-annotate the intra-document relations for *both* notes A and B above as *cancer* CON-SUB *adenocarcinoma*, which streamlined cross-document decisions – *cancer*$_A$ *IDENT cancer*$_B$ and *adenocarcinoma*$_A$ *IDENT adenocarcinoma*$_B$ – while preserving the "quasi-identical" relation (Hovy et al., 2013) between the cancer and adenocarcinoma terms. While this solution does not fully resolve the problem of binary annotation choices, it does provide more "wiggle room" along the meaning spectrum (Cruse, 1986) by introducing a third value – two mentions may be identical, non-identical, or mereologically (part-whole) related.

In keeping with our primary goal of enabling timeline extraction, we implemented CON-SUB as a TLINK since it conveys true temporal containment. CON-SUB, however, differs from other TLINKs, which were constrained by proximity and lexical cues, as discussed in section 2. The fact that CON-SUB also represents structural information allowed us to treat it like a coreference/bridging relation in terms of permissible textual evidence for link creation: namely, semantic scripts. These may be defined as: "a stereotypical sequence of events" (Araki et al., 2014) or "prototypical schematic sequences of events" (Chambers and Jurafsky, 2008). We can expect a surgery, for example, to consist of certain, typical subevents (incisions, subprocedures, anesthesia administration, etc.), which therefore enables annotators to look throughout the whole document for lexical items with meanings that fit those subevents. The concept of semantic scripts is what facilitates attainable long-distance coreference /bridging linking, and therefore, long-distance CON-SUB linking. This is obviously not the case for non-subevent CONTAINS relations.

We were therefore able to revise the THYME annotations to accommodate CON-SUB in two ways: First, we converted CONTAINS links to CON-SUB as appropriate, e.g.: *treatment CONTAINS radiation* became *treatment CON-SUB radiation*. Secondly, we added certain long-distance CON-SUB links for which there were no pre-existing CONTAINS annotations. These changes were made via a double-blind annotation process followed by an adjudication pass. The annotation team for the entire project (including the cross-document stage) consisted of nine annotators, eight of whom either had or were obtaining undergraduate or graduate degrees in linguistics. The ninth annotator was a physician who received on-the-job linguistics training and focused primarily on annotation subtasks that demanded considerable medical knowledge. Additionally, we consulted regularly with an oncologist and a medical coder with a decade of NLP annotation experience.

4.2 Inter-annotator Agreement

While the gold intra-document CON-SUB relations enabled high cross-document coreference agreement (93.77%), the IAA score for single-file CON-SUB links themselves was low at 34.14%[4]. Several factors contributed to this, but we focus on one major one here. The size and complexity of the guidelines[5] reflected the size and complexity of the task, augmenting the already heavy cognitive burden on annotators.

The cross-document task exposes greater nuance in event structure, involves greater variability of word sense, and attempts to join narratives that are temporally and linguistically disjunct. Annotation guidelines that set out to accurately represent information that is inherently nuanced, variable, and noncohesive will not be simple (see Savkov et al., 2016).

In determining guidelines for the subevent relation, we found that the best course for handling variability in lexical sense differed depending on the semantic potential (that is, how

[3] We did not use the pre-existing WHOLE-PART relation in order to preserve a distinction between events and entities, which also streamlined the annotation process. (WHOLE-PART was most often used to represent anatomical-site relations, such as *colon*$_{WHOLE}$ – *splenix flexure*$_{PART.}$)

[4] This score represents annotator-annotator agreement. We did not score annotator-gold, since adjudicators were permitted to make some specific changes that annotators were not.

[5] https://www.colorado.edu/lab/clear/projects/computationalsemantics/annotation

adaptable a word is to different meanings; see Evans, 2006) of the individual words used most frequently for each event category (i.e., semantic script)[6]. Not surprisingly, lexically-specific rules contributed significantly to the sheer size and complexity of the guidelines. Compare the following:

6. *We are seeing the patient for recent diagnosis of colon* ***cancer****. The* ***tumor*** *in her colon is quite large.*

7. *We recommended adjuvant* ***treatment****. Patient will return to start* ***chemo*** *in two weeks.*

Unmodified, *treatment* has an impoverished semantic potential (Evans, 2006); the meaning it conveys in itself is sparse, yielding a semantic flexibility that allows it to represent a wide range of referents. It is intuitive to understand it in (7) as coreferential with the much more precise *chemo*. *Cancer*, however, has a richer potential; it conveys a temporally-extensive disease that may have multiple manifestations (a primary tumor, recurrent tumors, metastatic tumors, etc.), rendering it less amenable to a coreferential link with *tumor*.

Therefore, for the cancer semantic script, annotators were asked to distinguish between terms that are defined more generally (*cancer*, *disease*) and terms that are more specific (*adenocarcinoma*, *tumor*, *metastasis*, *mass*, etc.), such that the specific terms were always subevents of the general terms, *regardless of pragmatic support for wide or narrow readings for a given term*. The reason for this has already been partially discussed in example (4); we add here that the semantic rigidity of *cancer* (compared to *treatment/therapy* terms, for example) also informed this choice.

This required a degree of abstraction from the text and an intentional suppression of instinctive linguistic judgments on the single-document level – for example, *adenocarcinoma* in (4) was re-annotated as a subevent of *cancer*, in spite of the fact that the text supports the wide reading.[7]

On the other hand, for the treatment/therapy semantic script, annotators were to determine relations more intuitively. Compare the resulting impact for cross-document linking in (8) to (4):

8. **Note A**: *We recommended adjuvant* ***treatment****. Patient will return for the first day of* ***chemo*** *in two weeks.*
   a. *treatment* IDENT *chemo*

Note B: *Adjuvant* ***treatment*** *consisted of four months of* ***chemo*** *and* ***radiation*** *and was without complication.*
   b. *treatment* CON-SUB *chemo*
   c. *treatment* CON-SUB *radiation*

If we linked *treatment*$_{A}$ to *treatment*$_{B}$ and *chemo*$_{A}$ to *chemo*$_{B}$, the product would be the same undesirable entailments discussed in (4): in this case, that radiation is a subevent of chemotherapy (an entirely different treatment), and that the chemo event temporally contains itself. Here, however, our solution differed: Rather than re-annotate (8a) as *treatment* CON-SUB *chemo*, we simply chose to leave *treatment*$_{A}$ and *treatment*$_{B}$ unlinked in cross-document annotation. The only cross-document link for this context was *chemo*$_{A}$ IDENT *chemo*$_{B}$. Again, this is due to the semantic malleability of *treatment.* Leaving two identical strings (*adjuvant treatment*) unlinked to each other is less problematic when that string regularly refers to a wide variety of referents. Furthermore, in experiments with re-analysis that paralleled the cancer semantic script approach, we found that attempting to always annotate *treatment* as an umbrella event proliferated unnecessary nested relations (because of how modifiable it is), increasing disagreement potential.

Unsurprisingly, an analysis of CON-SUB disagreements suggests that annotators struggled to remember when to abstract terms away from the context and when to interpret them intuitively; an example like (7) was apt to produce a disagreement, shown here in (9):

[6] Due to time constraints, we only added subevent relations for four categories of events: cancer, cancer treatment (surgeries, chemotherapy, and radiation), medications, and chronic diseases.

[7] In and of itself, this is not unreasonable given that the cross-document annotation task inherently lacks the linguistic cues supplied by a single, cohesive discourse (Wright-Bettner et al., 2019).

Relation type	Description	Link
CONTAINS supertype		
CONTAINS	M1 temporally contains M2	[M1] CONTAINS [M2]
CONTAINS-SUBEVENT	M1 temporally contains M2 and M2 is part of the structure of M1	[M1] CON-SUB [M2]
OVERLAP supertype		
OVERLAP	M1 temporally overlaps M2	[M1] OVERLAPS [M2]
NOTED-ON	M1 temporally overlaps and is observed on M2 *(used to link tests to test results)*	[M1] NOTED-ON [M2]
Other temporal links		
BEFORE	M1 temporally begins and ends before M2 begins	[M1] BEFORE [M2]
BEGINS-ON	The start of M1 begins at the end of M2	[M1] BEGINS-ON [M2]
ENDS-ON	The end of M1 ends at the start of M2	[M1] ENDS-ON [M2]

Table 1: Revised intra-document temporal links in the THYME colon cancer corpus

*9. We recommended adjuvant **treatment**. Patient will return to start **chemo** in two weeks.*

Annotator A: treatment IDENT chemo
Annotator B: treatment CON-SUB chemo

While Annotator A correctly interpreted the terms as coreferential (based on the context), Annotator B mistakenly followed the approach for the cancer semantic script in analyzing “treatment” as an umbrella event.

The annotation task was already demanding, requiring annotators to learn and assimilate specialized medical knowledge and terminology from the clinical texts, which themselves are written in heavy shorthand and with a mix of template language and free text that sometimes conflict. In addition, the non-linguistic nature of the cross-document component forced the creation of several counterintuitive annotation rules, which frequently (but not always) required them to ignore real linguistic cues.

Finally, we did not calculate IAA for non-subevent CONTAINS links because they already existed. They did, however, change slightly, along with all the TLINK types, since the auto-merger of the temporal and coreference annotations (discussed in section 2) produced some informational conflicts that we calibrated in a manual single-annotated pre-processing pass. However, as a proxy for annotator agreement, we show below that learnability improved for all temporal relations.

5 Summary of Refined THYME Relations

The pre-existing CONTAINS annotations were revised in part through the addition of two new links, both of which convey temporal and non-temporal information. The original CONTAINS and OVERLAP relations were therefore reimagined as supertypes[8], each consisting of two subtypes. All links are described in Table 1. We refer to these fine-grained THYME annotations as THYME+ and to the original THYME annotations as THYME.

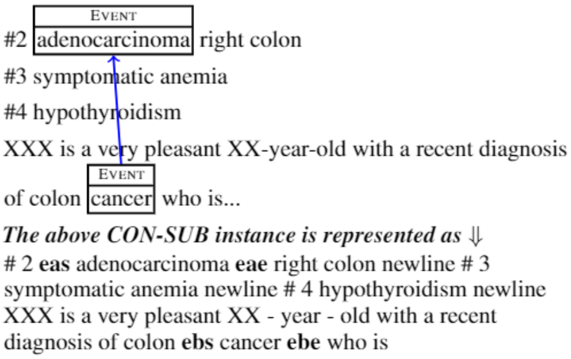
In the next section we demonstrate the learnability of the refined THYME+ temporal relations with state-of-the-art transformer methods. We report results that establish baselines for the THYME+ corpus for further methods development and offer insights into the challenges we faced which we view as exciting venues for future research.

6 Learning Refined Temporal Relations

Following the same window-based processing (using a span of contiguous tokens disregarding sentence boundaries for generating relational candidates) and argument-marking mechanism developed by the prior study that achieved the state-of–the-art results on THYME (Lin et al.,

[8] It is worth noting that while CONTAINS itself may be thought of as a specific subtype of an overlap temporal relation, we use the OVERLAP category specifically for non-containment temporal overlap or underspecified cases for which we cannot claim containment.

2019), we tested a series of pre-trained models for extracting both within and cross-sentence temporal relations (i.e. TLINKs) in a multi-class classification fashion. Figure 1 shows a CON-SUB relation between "cancer" and "adenocarcinoma", and its representation as a token sequence. Special token pairs, "eas" and "eae", "ebs" and "ebe" mark the events of interest in the sequence.


EVENT
#2 adenocarcinoma right colon
#3 symptomatic anemia
#4 hypothyroidism
XXX is a very pleasant XX-year-old with a recent diagnosis
EVENT
of colon cancer who is...

The above CON-SUB instance is represented as ⇓
2 **eas** adenocarcinoma **eae** right colon newline # 3 symptomatic anemia newline # 4 hypothyroidism newline XXX is a very pleasant XX - year - old with a recent diagnosis of colon **ebs** cancer **ebe** who is

Figure 1: CON-SUB instance and its representation as a token sequence

Pre-trained models BERT (Devlin et al., 2019), BioBERT (Lee et al., 2020), Xlnet (Yang et al., 2019), ALBERT (Lan et al., 2020), RoBERTa (Liu et al., 2019), BART (Lewis et al., 2019), and SpanBERT (Joshi et al., 2020) were used to encode each input sequence with a relational candidate by the [CLS] token, which was fed to the classification layer to predict the relation type for every relational pair candidate. For some of the popular models, such as BERT, RoBERTa, and BART, we also tried their *large* version in addition to their *base* releases.

We used NVIDIA GTX Titan Xp GPU and Titan RTX GPU cluster of 7 nodes for fine-tuning the pre-trained models. The fine-tuning is done with HuggingFace's Transformers API (Wolf et al., 2019) and the TensorFlow-based BERT API, with batch size selected from (16, 32), a 60-token sliding window for generating candidate relational pairs, a maximal sequence length of 100 word pieces to accommodate all word pieces from the 60 tokens, and a learning rate selected from (1e-5, 2e-5, 3e-5, 5e-5). The performance was evaluated by the standard Clinical TempEval (Bethard et al., 2017) evaluation script, modified only to accommodate the new categories.

7 Experimental Results

The model that performed best on THYME (BioBERT-base) was trained and evaluated on THYME+ annotations. The first two rows of Table 2 gauge performance purely based on the refinements of the THYME+ annotations. Splitting CONTAINS into CONTAINS and CON-SUB relations and OVERLAP into OVERLAP and NOTED-ON relations leads to better learnability: CONTAINS goes from 0.664 F1 on THYME to 0.748 F1 on THYME+, and OVERLAP goes from 0.179 on THYME to 0.416 on THYME+. The best results for the new categories of CON-SUB and NOTED-ON are 0.072 F1 and 0.744 F1 respectively – results that establish baselines for these two new temporal relations. The performance on all types of relations for THYME+ is 0.625 F1 compared to 0.548 for THYME (Table 2, *Overall* column, rows 1 and 2).

Lin et al., 2019 report 0.684 F1 for THYME CONTAINS, however the result is achieved when training on and evaluating for only the CONTAINS links, and augmenting the training data with automatically generated CONTAINS relations. Thus, it is not a fair comparison to use for the results reported in Table 2.

Of the models beyond BioBERT that we explored, BART-large was the most successful. The result with BART-large was 0.748 F1 (Table 2, *CONTAINS* column, row 3). In general, certain pre-trained models, like BioBERT and BART, yield better results than the other models. BioBERT is pre-trained on biomedical text and thus can help encode clinical text better. BART masks a contiguous span of text rather than random tokens, which can be helpful for encoding clinical text where many event and temporal expressions consist of multiples tokens, e.g. "ascending colon cancer".

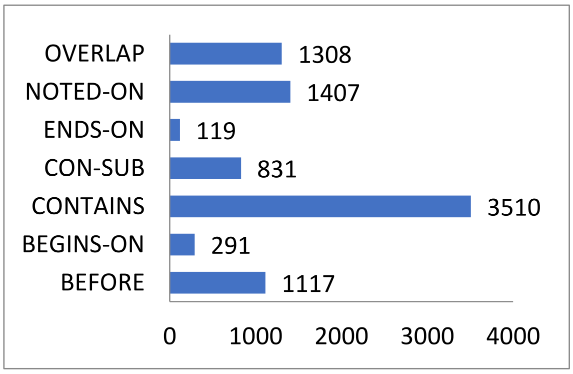


Figure 2: Number of instances for each temporal relation type in the colon test set.

Model	*Data*	*BEFORE*			*BEGINS-ON*			*CONTAINS*			*CON-SUB*		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
BioBERT	THYME	0.168	0.311	0.218	0.263	0.070	0.110	0.701	0.631	0.664			
BioBERT	THYME+	0.278	**0.458**	0.346	**0.423**	0.175	**0.248**	0.793	0.708	0.748			
BART	THYME+	**0.313**	0.434	**0.364**	0.383	**0.179**	0.244	0.791	0.709	0.748	0.375	0.040	0.072
BART-	THYME+	0.300	0.422	0.351	0.378	0.175	0.239	**0.796**	**0.710**	**0.750**			

Model	*Data*	*ENDS-ON*			*NOTED-ON*			*OVERLAP*			*OVERALL*		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
BioBERT	THYME	**0.194**	0.099	0.131				0.174	0.185	0.179	0.594	0.508	0.548
BioBERT	THYME+	0.112	**0.210**	0.146	**0.786**	0.706	**0.744**	0.353	**0.508**	0.416	0.696	**0.568**	0.625
BART	THYME+	0.120	**0.210**	0.153	0.787	0.702	0.742	0.401	0.482	**0.438**	0.713	0.567	**0.632**
BART-	THYME+	0.124	**0.210**	**0.156**	**0.786**	**0.707**	**0.744**	**0.404**	0.470	0.435	**0.718**	0.558	0.628

Table 2: Model performance on THYME and THYME+.
BART = BART large trained on all THYME+ links.
BART- = BART large trained on all THYME+ links excluding CON-SUB

CONTAINS and NOTED-ON are the two types of temporal relations that are overwhelmingly the most frequent in THYME+ (Figure 2), thus of the greatest interest from an applications point of view. THYME+ CONTAINS best result of 0.75 F1 and THYME+ NOTED-ON best result of 0.744 F1 are higher than the state-of-the-art THYME result on CONTAINS of 0.684 F1 (Lin et al., 2019). BEGINS-ON and ENDS-ON categories have the fewest instances, as shown in Figure 2, so their low performance could thus be due to their lack of representation. CON-SUB's low performance could be attributed to their coreferential nature and the long distance relations between the two arguments which in many cases surpass our 60-token window limit.

Table 2, row 4 presents the results with the BART-large model trained and evaluated when excluding CON-SUB links. The 0.750 F1 on CONTAINS is similar to the 0.748 F1 on CONTAINS when training and evaluating on all THYME+ relations. Thus, while the model is not able to accurately predict CON-SUB relations, including them does not appear to cause confusion for the model.

8 Discussion

Splitting CONTAINS into CONTAINS and CON-SUB categories improved the annotation quality of the CONTAINS class instances in THYME+. The CONTAINS class is the most frequent relation in clinical text and very easy for transitive closure to operate upon. As we already pointed out, the CONTAINS performance on THYME+ is improved from 0.664 F1 to 0.748 F1 using the same BioBERT model (Table 2, *CONTAINS* column, row 1 and 1), with both improved P and R.

The creation of gold NOTED-ON instances was straightforward, thus with high quality. NOTED-ON is the second most frequent relation in the corpus. 65.12% of the NOTED-ON relations are within one sentence and very few cases are long-distance. This makes the NOTED-ON class very learnable.

The results on THYME+ BEFORE, BEGINS-ON, ENDS-ON, and OVERLAP also improved compared to their respective THYME results. We attribute it to the improved performance of CONTAINS and NOTED-ON links as the definitions, hence the space separation, are tightened.

An error analysis of the CON-SUB relations showed that the main error consisted of missed links that relied on the semantic-script concept discussed in section 4.2. For example, the system often failed to capture the CON-SUB relation between *cancer* and *adenocarcinoma*. These are often long-distance relations: of all gold CON-SUB relations, 67.63% of them are beyond our 60-token window limit. Even if we focused on those within-window CON-SUB relations (32.37% of total), the performance was still low (0.441 P, 0.112 R, 0.178 F1), which showed our models had not captured the peculiarities of the CON-SUB class. The fact that the majority of instances of CON-SUB class are long-distance is quite different from the other TLINKs and hard for transitive closure to act upon, suggesting they might need a different approach than the other TLINKs.

However, one error category that could be resolved by transitive closure are links the system marked that are not present locally in the gold, but are correct by inference via the coreference relations. Consider:

10. ***Procedure** to include hernia **repair**.*
a. System: *Procedure* CON-SUB *repair*

b. Gold: No local TLINK.

In this example, one or both terms are linked as IDENT to earlier mentions in the note. To save time and visual clutter, annotators only linked bridging relations like CON-SUB to first mentions, under the assumption that redundant information could be retrieved from the IDENT chains. Therefore, there is no error here if the IDENT chains are taken into account.

In short, CON-SUB relations capture temporality and mereological relations. Thus representations and methods for combined temporality and coreference are suitable venues to explore.

Except for CONTAINS, the other types of TLINKs have relatively low numbers of instances. The creation of instances for the low number of temporal relation types is limited by two main factors: (1) availability of data due to privacy constraints on EMR clinical narratives, and (2) the time, effort and budget required for such an activity. We have shown that with enough training instances (see CONTAINS), temporal relations are learnable at improved rates with the latest state-of-the-art methods. Although NOTED-ON has a similarly low number of instances as OVERLAP and BEFORE, it is highly learnable (0.744 F1) which we attribute to its semantic-script characteristics as discussed in section 4.2. This suggests that there are several paths to explore among which are: (1) re-defining and refining the other types of relations, and (2) devising methods for relations with low number of instances.

9 Conclusion

In this study, we presented our refinements for temporal relation annotations of the THYME corpus resulting in the THYME+ corpus. The main modifications are in re-defining CONTAINS and OVERLAP relations into CONTAINS, CONTAINS-SUBEVENT, OVERLAP and NOTED-ON. This strategy is theoretically based and led to better learnability with the latest transformer methods. Our results establish baselines for future methods -- CONTAINS 0.750 F1 OVERLAP 0.438 F1, CON-SUB 0.072 F1 and NOTED-ON 0.744 F1.

Acknowledgements

The work was supported by funding from the United States National Institutes of Health -- R01LM010090 from the National Library of Medicine and U24CA248010 and UG3CA243120 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

We would like to thank: Samuel Beer, Hayley Coniglio, Jameson Ducey, Ahmed Elsayed, Dana Green, Matthew Oh, Michael Regan, and Adam Wiemerslage for annotation; Piet deGroen and David Harris for medical advice; and Wei-Te Chen and Skatje Myers for technical support.

References

Jun Araki, Zhengzhong Liu, Eduard Hovy, and Teruko Mitamura. 2014. Detecting Subevent Structure for Event Coreference Resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, page 6, Reykjavik, Iceland, May.

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

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised Learning of Narrative Event Chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio, June. Association for Computational Linguistics.

David Cruse. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge, UK.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May. arXiv: 1810.04805.

Vyvyan Evans. 2006. Lexical concepts, cognitive models and meaning-construction. *Cognitive Linguistics*, 17(4):491–534.

Jerry R. Hobbs. 1985. Granularity. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, volume 1, pages 432–435. August.

Eduard Hovy, Teruko Mitamura, Felisa Verdejo, Jun Araki, and Andrew Philpot. 2013. Events are Not Simple: Identity, Non-Identity, and Quasi-Identity. In *Workshop on Events: Definition, Detection, Coreference, and Representation*, pages 21–28, Atlanta, Georgia, June. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *arXiv:1907.10529 [cs]*, January. arXiv: 1907.10529.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv:1909.11942 [cs]*, February. arXiv: 1909.11942.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, February.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv:1910.13461 [cs, stat]*, October. arXiv: 1910.13461.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2019. A BERT-based universal model for both within-and cross-sentence clinical temporal relation extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 65–71, Minneapolis, MN.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*, July. arXiv: 1907.11692.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas, November. Association for Computational Linguistics.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An International Standard for Semantic Annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, page 4, Valletta, Malta, May.

Preethi Raghavan, Eric Fosler-Lussier, Noémie Elhadad, and Albert M. Lai. 2014. Cross-narrative Temporal Ordering of Medical Events. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 998–1008, Baltimore, Maryland, June. Association for Computational Linguistics.

Marta Recasens, Eduard Hovy, and M. Antònia Martí. 2011. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152, May.

Aleksandar Savkov, John Carroll, Rob Koeling, and Jackie Cassell. 2016. Annotating patient clinical records with syntactic chunks and named entities: the Harvey Corpus. *Language Resources and Evaluation*, 50:523–548.

Barry Smith, Werner Ceusters, Bert Klagges, Jacob Köhler, Anand Kumar, Jane Lomax, Chris Mungall, Fabian Neuhaus, Alan L. Rector, and Cornelius Rosse. 2005. Relations in biomedical ontologies. *Genome Biology*, 6(5):R46, April.

Zhiyi Song, Ann Bies, Justin Mott, Xuansong Li, Stephanie Strassel, and Christopher Caruso. 2018. Cross-Document, Cross-Language Event Coreference Annotation Using Event Hoppers. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, et al. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv e-prints*, 1910:arXiv:1910.03771, October.

Kristin Wright-Bettner, Martha Palmer, Guergana Savova, Piet de Groen, and Timothy Miller. 2019. Cross-document coreference: An approach to capturing coreference without context. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 1–10, Hong Kong, November. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.