Empirical Exploration of the Challenges in Temporal Relation Extraction from Clinical Text

Diana Galvan-Sosa[†], Koji Matsuda^{††,†}, Naoaki Okazaki^{†††} and Kentaro Inui^{†,††}

Time is an important concept in human-cognition, fundamental to a wide range of reasoning tasks in the clinical domain. Results of the Clinical TempEval 2016 challenge, a set of shared tasks that evaluate temporal information extraction systems in the clinical domain, indicate that current state-of-the-art systems do well in solving event and time expression identification but perform poorly in temporal relation extraction. This study aims to identify and analyze the reason(s) for this uneven performance. It adapts a general domain tree-based bidirectional long short-term memory recurrent neural network model for semantic relation extraction to the task of temporal relation extraction in the clinical domain, and tests the system in a binary and multi-class classification setting by experimenting with general and in-domain word embeddings. Its results outperform the best Clinical TempEval 2016 system and the current state-of-the-art model. However, there is still a significant gap between the system and human performance. Consequently, this study delivers a deep analysis of the results, identifying a high incidence of nouns as events and class overlapping as posing major challenges in this task.

Key Words: *Clinical Text, Natural Language Processing, Temporal Relation Extraction, Aspect, Nominal Events*

1 Introduction

Human reasoning has to do with time. High-level cognition concepts, such as *duration* and *sequence*, influence the structure of human interaction with the external world. Temporal reasoning is a fundamental ability not only in humans but also in intelligent systems. In Natural Language Processing (NLP), Temporal Information Extraction (TIE) is an active research area where the ultimate goal is to be able to represent the development of a story over time. This is key to text processing tasks including question answering (UzZaman, Llorens, and Allen 2012) and text summarization (Jung, Allen, Blaylock, de Beaumont, Galescu, and Swift 2011), and it follows the traditional pipeline of named entity recognition and relation extraction separately.

[†] Graduate School of Information Sciences, Tohoku University

^{††} RIKEN Center for Advanced Intelligence Project

^{†††} School of Computing, Tokyo Institute of Technology

In a temporal context, entities are typically classified as either events or time expressions and temporal relations describe how their time intervals interact, assuming a linear model of time. Besides reasoning, choosing an accurate representation of time is challenging. In language, events are typically conceptualized as something that occurs, and they all have some duration. In the clinical domain, events can range from procedures to diseases to diagnoses, or to anything that the patient experiences. For simplicity, a point-based temporal logic is typically used to associate two periods of time. For example, given an event A (“surgery”) and a time expression B (“tomorrow”), where A precedes B, we can infer a temporal relation BEFORE between A and B. Intuitively, we can also say that B comes AFTER A. The main problem with this temporal logic is that several temporal relations, which are not necessarily relevant to the reason about the situation described, can be identified within a text. Narrative containers are defined by Pustejovsky and Stubbs (2011) as an effort to reduce the scope of temporal relations between pairs of events and time expressions. As illustrated in Figure 1, narrative containers can be thought of as temporal buckets in which an event or a series of events may fall (Styler IV, Bethard, Finan, Palmer, Pradhan, de Groen, Erickson, Miller, Lin, Savova, and Pustejovsky 2014). They help visualize the temporal relations within a text and facilitate the identification of other temporal

Fig. 1 Example temporal relation annotation with and without using narrative containers.

relation types. Until now, the only corpus annotated with this schema is limited to clinical texts.

1.1 Clinical TempEval Challenges

Research on TIE has been instigated by Temporal Evaluation (TempEval) shared tasks that are focused on processing news article documents (Verhagen, Gaizauskas, Schilder, Hepple, Katz, and Pustejovsky 2007; Verhagen, Saurí, Caselli, and Pustejovsky 2010; UzZaman, Llorens, Derczynski, Allen, Verhagen, and Pustejovsky 2013). However, in recent years, due to the high role of temporal reasoning in the interpretation of clinical narratives, the target domain has been shifted to the clinical domain. The resulting Clinical TempEval challenges (Bethard, Derczynski, Savova, Pustejovsky, and Verhagen 2015; Bethard, Savova, Chen, Derczynski, Pustejovsky, and Verhagen 2016; Bethard, Savova, Palmer, and Pustejovsky 2017) evaluate systems on temporal information extraction from clinical notes and pathology reports from colon cancer patients, defining a series of sub-tasks that aim to identify temporal entities (EVENTS and TIMEX3: time expressions) and the temporal relations (TLINK) between them. Participating systems can choose to use raw text as input (phase 1) or they can use raw text with EVENT and TIMEX3 annotations (phase 2), in which case their task is to only identify temporal relations. The temporal relation extraction track is further divided into two sub-tasks: (1) the identification of relations between events and the document creation time and (2) the identification of narrative container relations (TLINK:CONTAINS) between a directed pair (e_1, e_2) . In this case, e_1 and e_2 are entities of either EVENT or TIMEX3 type. Clinical TempEval 2017 (Bethard et al. 2017) introduced a new aspect to the challenge, which still maintains the aforementioned sub-tasks but sets a new goal—to explore how well the systems trained in one medical domain perform on data from another. Such systems are trained on colon cancer data but are instead tested on brain cancer data.

Results of the systems participating in Clinical TempEval 2016 suggest that they perform well on time-entity identification tasks. Nevertheless, temporal relation extraction has proven to be the most difficult task. UTHHealth (Lee, Xu, Wang, Zhang, Moon, Xu, and Wu 2016), the best ranked system in Clinical TempEval 2016, showed a significant gap of 0.25 when compared to human performance,¹ even when gold-standard entity annotations were provided. The improved task performance of recent works by Lin, Miller, Dligach, Bethard, and Savova (2016) and Leeuwenberg and Moens (2017) further enhanced the credibility of UTHHealth’s results, but the gap with respect to humans is still around 0.21. Regardless of the increase in the annotation agreement of temporal

¹ There are two scores for human performance: inter-annotator agreement and annotator-adjudicator agreement. We consider ann-adj as the upper bound performance since the models are trained on the adjudicated data, not on the individual annotator data (Bethard et al. 2015, 2016, 2017)

relations by relying on narrative containers, there is a consensus within the research community regarding the difficulties experienced in TIE. However, the reasons behind the skewed results between entity and temporal relation predictions still remain unclear.

1.2 Previous Work

Until Clinical TempEval 2016, classic machine learning algorithms for classification such as conditional random fields, support vector machines (SVM) and logistic regression with a variety of features (e.g., lexical, syntactic, and morphological) were the predominant approach to TIE (Bethard et al. 2015, 2016). In fact, the best performance was achieved by the UTHealth team (Lee et al. 2016) using an end-to-end system based on a linear and structural Hidden Markov Model (HMM)-SVM. Only a few teams tried a neural based method, including recurrent neural networks-based (RNN) models (Fries 2016) and convolutional neural networks-based (CNN) models (Chikka 2016; Li and Huang 2016). Furthermore, among those teams, only Chikka (2016) participated in the CONTAINS identification task, being around 0.30 below UTHealth’s top performance.

Recent works by Lin et al. (2016), Dligach, Miller, Lin, Bethard, and Savova (2017) and Leeuwenberg and Moens (2017) followed the settings of Clinical TempEval 2016 but they did not participate in the competition. Even though Leeuwenberg and Moens (2017) developed a new state-of-the-art model for temporal relation extraction, their results are still below human performance. Moreover, none of the aforementioned works provide a detailed discussion of *why* the current performance is so low and *how* the results on temporal relation extraction can be improved, save for Leeuwenberg and Moens, who in their first attempt on tackling this task on Clinical TempEval 2016 (Leeuwenberg and Moens 2016), identified false negatives as their major problem.

Rather than a model’s architecture or a dataset size, we believe that the complexity of temporal representation in natural language is likely to be the main cause of the low performance on temporal relation. *Tense* and *aspect* are the two grammatical means of expressing the notion of time in English, but little has been discussed about the latter in clinical texts. Furthermore, the focus of previous work on temporal relation extraction is set on narrative containers, relegating the identification of other temporal relations to a second place. However, we believe that the key is to look at the whole set of temporal types to achieve the ultimate goal of developing systems that can reason about time to automatically create a timeline of a patient’s health care.

This study contributes to the current understanding of how temporal relations work in the clinical domain. It begins by illustrating how we adapt a general domain neural model for

semantic relation extraction to temporal relation extraction from clinical text. It then analyzes the adopted system’s overall performance, including the identification of narrative containers. Next, it discusses two major problems encountered when working with the narrative container’s annotation schema, ending with a discussion on the necessary efforts needed to further improve the performance of the current state-of-the-art temporal relation extraction systems to perform on par with humans in terms of efficiency in completing the same tasks.

2 Methods

2.1 From Relation Extraction to Temporal Relation Extraction

To determine the challenges in temporal relation extraction from clinical text, this study adapts a general domain relation extraction model. In NLP literature, the term “relation extraction” is a short form for “semantic relation extraction”, which is the existing association between the meaning of words, phrases, or sentences. Time concepts, such as duration and sequence, are embedded in a word’s meaning (e.g., “bleeding” is an action that usually lasts for a moment and comes after another action like “cutting”). Therefore, semantic relations and temporal relations naturally overlap.

Relation extraction is a well-studied task in NLP, where besides semantics, word sequence structures—such as recurrent neural networks (RNN) and linguistic features like the path of target words in the dependency tree—have shown to be effective (Xu, Feng, Huang, and Zhao 2015). There is already a relation extraction model that integrates all of these elements—the end-to-end tree-based bidirectional long short-term memory-RNN model of Miwa and Bansal (2016). Due to its availability and state-of-the-art performance, we chose this model over Leeuwenberg and Moens’s (2017) system. Moreover, we aim to take advantage of word sequence and dependency tree structures to further improve the performance on the Clinical TempEval relation extraction task. Given a sentence, Miwa and Bansal’s (2016) three-layer model (i.e., embedding, sequence and dependency layers) jointly identifies entities and the relations between them. The model receives a sentence and an annotation file with a pair of terms as input and outputs the predicted relation type and directionality of the terms: (t_1, t_2) if t_1 is the source and t_2 the target, and (t_2, t_1) otherwise.

2.2 Experimental Settings

Similar to Clinical TempEval 2016, we used the THYME corpus (Styler IV et al. 2014), a dataset of 600 clinical notes and pathology reports from colon cancer patients at the Mayo Clinic.

The corpus is annotated at the document level and the identified entities are given a set of attributes depending on their type (i.e., *DocTimeRel*, *Type*, *Polarity*, *Degree*, *Contextual Modality*, and *Contextual Aspect* for EVENTS and *Class* for TIMEX3). Temporal relation annotations specify source and target entities along with one of the following TLINK types: BEFORE, BEGINS-ON, CONTAINS, ENDS-ON, and OVERLAP. Considering Miwa and Bansal’s (2016) processes one sentence at a time, data was pre-processed to get sentence-level annotations. Any two EVENT/TIMEX3 can be a candidate pair. Therefore, all the entities in a sentence were used in generating all pair permutations as candidates. Pairs that did not have any temporal relations were then labeled as NONE (see Appendix A.1). The frequency of TLINKS in the THYME corpus is higher than the relations in the SemEval-2010 Task 8 dataset (Hendrickx, Kim, Kozareva, Nakov, Ó Séaghdha, Padó, Pennacchiotti, Romano, and Szpakowicz 2010), on which Miwa and Bansal’s (2016) model was tested for relation classification. For this reason, we did not consider it necessary to extend the set of TLINKS to its transitive closure for data augmentation (i.e., $A \text{ CONTAINS } B \wedge B \text{ CONTAINS } C \rightarrow A \text{ CONTAINS } C$). Table 1 and Table 2 detail the resulting datasets.

In addition to the model’s default Wikipedia word embeddings, we trained word vectors of 200 dimensions using word2vec (Mikolov, Chen, Corrado, and Dean 2013) on a subset of PubMed2014.² PubMed2014 has 10,969,353 abstracts from 1,118,934 different journals. From those, we selected 634,813 abstracts from 38,677 journals related to Oncology and Gastroenterology. The MIMIC II clinical corpus (Saeed, Villarroel, Reisner, Clifford, Lehman, Moody, Heldt, Kyaw, Moody, and Mark 2011) is closer in genre to the THYME dataset, but due to its nature, one must get an application approval for its use. For simplicity purposes, we instead chose PubMed. Next, we conducted four experiments at the intra-sentential level. The first experiment

Table 1 Label distribution of pre-processed dataset for binary classification

TLINK	Train	Test	Dev
CONTAINS	8,653	4,554	4,780
NONE	43,643	20,465	24,046
Total	52,296	25,019	28,826

Table 2 Label distribution of pre-processed dataset for multi-class classification

TLINK	Train	Test	Dev
BEFORE	1,839	982	917
BEGINS-ON	717	363	298
CONTAINS	8,653	4,554	4,780
ENDS-ON	334	138	151
OVERLAP	2,388	1,186	1,582
NONE	43,643	20,465	24,046
Total	57,574	27,688	31,774

² https://www.nlm.nih.gov/databases/download/pubmed_medline.html

followed the Clinical TempEval 2016, focusing only on the identification of the CONTAINS type. The remaining experiments included all of the five annotated TLINKS. Further details of each of the experiments are given below:

- (I) **TLINK:CONTAINS binary classification:** In order to obtain results comparable to Lee et al. (2016), the best ranked system in Clinical TempEval 2016, we only considered TLINK:CONTAINS instances. The model chooses between CONTAINS and NONE relations.
- (II) **Multi-class classification with Wikipedia word embeddings:** To test the model in a real-world setting (i.e., a document that not only includes CONTAINS relations), we added the remaining pairs in the gold standard that have any of the other TLINK types to the train and test sets.
- (III) **Multi-class classification with PubMed word embeddings:** In addition to the previous setting (II), we used word embeddings trained on a subset of PubMed instead of the default word vectors trained on Wikipedia.
- (IV) **Multi-class classification with PubMed word embeddings and filtered negative examples:** Two extra difficulties of temporal relations are to ascertain whether an EVENT happened or not in a clinical context, and evaluate whether the said EVENT actually relates to the patient. For this reason, the THYME corpus differentiates “real” (*Contextual Modality*: ACTUAL or HEDGED) from “non-real” (*Contextual Modality*: HYPOTHETICAL or GENERIC) events. Real events cannot be related to non-real events. Therefore, in addition to the previous setting (III), we experimented removing a candidate pair whenever the e_1 contextual modality value³ was ACTUAL or HEDGED and e_2 had HYPOTHETICAL or GENERIC modality, and vice versa.

3 Results

The evaluations were performed using the official Clinical TempEval scorer.⁴ Table 3 shows performance on the CONTAINS identification task as a binary classification problem. The first row shows the top performance in Clinical TempEval 2016, while the second row is a result outside of the competition. We obtained an F1 score of 0.633, outperforming both UTHealth and Lin et al. (2016). Our model shows a high precision, but a lower recall than UTHealth; this can be attributed to the NONE relations prevalent in the dataset. Despite the recorded improvement,

³ Note that entity attributes introduced at the beginning of this section were only used for pre-processing, and not as features in our model.

⁴ <http://alt.qcri.org/semEval2016/task12/index.php?id=software>

Table 3 Performance of systems and humans on identifying CONTAINS relations. Our results come from five different random seeds

System	P	R	F1
Lee et al., 2016 (UTHealth)	0.588	0.559	0.573
Lin et al., 2016	0.669	0.534	0.594
Our model	0.986	0.467	0.633
Human performance	—	—	0.817

Table 4 Results of the three multi-class classification experiments and Leeuwenberg and Moens’s (2017) Structured Perceptron (SP) best results on the THYME test set

TLINK	Multi-class classification									
	<i>Wikipedia word emb</i>			<i>PubMed word emb</i>			<i>PubMed word emb + FNE</i>			<i>SP</i>
	P	R	F1	P	R	F1	P	R	F1	F1
BEFORE	0.696	0.183	0.289	0.704	0.196	0.306	0.683	0.213	0.324	0.294
BEGINS-ON	0.628	0.082	0.145	0.620	0.110	0.186	0.635	0.114	0.194	0.159
CONTAINS	0.907	0.468	0.617	0.904	0.471	0.619	0.900	0.472	0.618	0.608
ENDS-ON	0.525	0.093	0.157	0.656	0.122	0.204	0.637	0.115	0.194	0.236
OVERLAP	0.494	0.121	0.195	0.526	0.124	0.201	0.518	0.131	0.209	0.204
Macro-F1	0.281			0.303			0.308			0.300

The SP results were reproduced from the original paper. The results come from five different random seeds. FNE – filtered negative examples.

it is not possible to compare our system’s performance with the current state-of-the-art set by Leeuwenberg and Moens’s (2017), which was obtained using a multi-class classification approach. Table 4 reports our experimental results with the three multi-class classification settings presented in Section 2.2. Switching from binary classification to multi-class classification, we observe a significant drop in precision and a lower F1 score. This is expected because the classifier now has more TLINKs as options to choose from. Despite this change, our model outperforms both UTHealth and the state-of-the-art model in terms of the F1 score of CONTAINS. Using PubMed word embeddings yielded the best F1 score for ENDS-ON and CONTAINS, and down-sampling negative examples on this setting improved the F1 score of BEFORE, ENDS-ON and OVERLAP. More details on the impact of using in-domain word embeddings and the FNE strategy are provided in Appendix B.

4 Discussions

Since this study did not change the architecture of Miwa and Bansal’s (2016) model, the

reader can, therefore, consult their study (Miwa and Bansal 2016) for a detailed discussion on the system’s performance on relation extraction. This section complements their discussion, which focuses on the linguistic characteristics of the dataset (clinical and temporal) that harms the system’s performance.

4.1 Error Analysis

Our error analysis focused on one-fourth of our experiments. Systems participating in the Clinical TempEval narrative container identification task only received credit for a pair of entities that they correctly identified the source, target, and the CONTAINS relation between them. Given this setting, we understand that even when using manual event and time annotations, the challenge is not only to predict the TLINK type but also the correct directionality of the entities. Therefore, part of our analysis aims to ascertain whether type classification or directionality identification is the most difficult task or if they are both equally problematic for the model. For this reason, we designed the confusion matrix using Miwa and Bansal’s (2016) output instead of the Clinical TempEval 2016 script output. The confusion matrix on Figure 2 shows

Fig. 2 Confusion matrix of our multi-class classification model with PubMed word embeddings on the dev set.

the results on the development set. Overall, due to the high number of negative instances, most of the false positives fall into the *None*(e_1, e_2) category. This type of relation is the reason why the system shows high precision. Apart from this, we can identify the performance on OVERLAP as our system's main problem. The accuracy in both *Overlap*(e_1, e_2) and *Overlap*(e_2, e_1) is considerably low, with the latter being the lowest among all types (with 0.021). Not even the performance on *BeginsOn*(e_2, e_1)—with 0.14—is as low as *Overlap*(e_2, e_1), although they have a similar number of instances (430 and 557, respectively). *Overlap*(e_1, e_2), with 0.14, is comparable to *BeginsOn*(e_2, e_1), despite having four times more instances (1,831 vs. 430). This explains why we focused our error analysis on OVERLAP. From Figure 2, we can observe that *Overlap*(e_1, e_2) is usually predicted as *Contains*(e_1, e_2) and *Overlap*(e_2, e_1) is predicted as *Contains*(e_2, e_1). In both cases, the directionality of the entities was correct but the system failed to identify the appropriate temporal relation. For *Overlap*(e_1, e_2), there were 112 sentences misclassified as *Contains*(e_1, e_2), while in *Overlap*(e_2, e_1) there were 32 *Contains*(e_2, e_1) misclassifications. EVENT-EVENT pairs were the predominant type of pair in the former while TIMEX3-EVENT was for the latter, with 101 and 25 instances, respectively. We took all of the aforementioned misclassified sentences for supplementary examination and discuss the reason(s) for these errors in the following section.

4.2 Temporal Relations and Aspectual Classes

Before proceeding further, it is important to understand the definition of OVERLAP and CONTAINS. Both temporal relations are closely related because they encompass the notion of two things happening at the same time. However, CONTAINS relations imply that the contained event (i.e., the target) occurs entirely within the temporal bounds of the event it is contained within (i.e., the source) while OVERLAP relations are those where containment is not entirely sure. Also, since e_1 OVERLAP e_2 means the same as e_2 OVERLAP e_1 , OVERLAP is the only symmetrical TLINK type.

4.2.1 Time representation: Interval algebra and linguistics

Strictly speaking, every entity occupies time. An entity's time interval is crucial in understanding its temporal relation with respect to another entity, especially in the case of CONTAINS and OVERLAP relations where the end point of the target is key in determining whether there is complete containment or not. The temporal relations used by the THYME project rely on Allen's (1990) interval algebra, a precise way of expressing time periods using clear start and end points. By comparing those, we can easily indicate the position of two events on the timeline.

However, the concept of time is widely discussed across disciplines and Allen’s representation is just one among many others. In linguistics, the expression of time is understood because of two important grammatical systems: *tense* and *aspect*. **Tense** is used to locate the time of an event being talked about with respect to the time at which the speaker utters the sentence (i.e., speech time), while **aspect** is used to describe how a speaker views the contour of a situation (i.e., as beginning, continuation, or completion), independent of which position in time this situation occupies (Li and Shirai 2000; Klein 2013). Therefore, when discussing a situation such as a patient having a surgery, we would use **past tense** if the surgery happened before speech time (*The patient had a surgery*), **future tense** if the surgery is about to happen (*The patient will have a surgery*), and **present tense** if the time of the surgery overlaps with the speech time (*The patient has a surgery*). Aspect, on the other hand, gives information about the surgery. The past tense in *The patient had a surgery* not only locates the surgery event before the speech time, but also conveys that the surgery was completed. This leads us to an important characteristic of tense and aspect—the boundaries between them are often not clear-cut (Li and Shirai 2000). The linguistic forms that express each of these notions tend to grammaticize into other categories (Bybee, Perkins, and Pagliuca 1994). In the case of English, the past tense form indicates past tense and perfective aspect (i.e., when a situation is reported in its entirety) simultaneously.

Having identified the functions of tense and aspect, we now focus on the latter. The study of aspect is commonly divided into the *grammatical aspect* (also known as viewpoint aspect (Smith 1983)) and the *lexical aspect* (also known as the situational aspect).⁵ Since temporal relations between events in the THYME project are thought in terms of their **start and endpoints**, the definition of the lexical aspect, which designates the internal temporal organization of the situation described by a verb, is particularly important to us (Klein 2013). One of the best known and widely accepted aspect classification is that of Vendler, which distinguishes four categories of the inherent semantics of verb and verb phrases: *activities*, *accomplishments*, *achievements*, and *states* (Vendler 1957). Figure 3 presents Vendler’s classification using Andersen’s (1990) schematization.

4.2.2 Aspectual classification of temporal entities

We expect that by categorizing the source and target entities of a relation as one of Vendler’s types, the underlying reasoning for the TLINK classification will be simplified. For example, categories with **no clear end points** (such as *activities* and *states*) are more likely to overlap with

⁵ For our discussion, we will use the term “aspect” to refer to lexical aspect.

those with **clear end points** (such as *accomplishments* and *achievements*). Figure 4 illustrates an OVERLAP and CONTAINS relations using Allen and Vendler’s representation of time periods. While analyzing OVERLAP relations that were mistaken for CONTAINS, we realized that only a few events were verbs. EVENTS in sentences 1, 3 and 9 in Table 5 are some examples of this (“invades,” “seeking,” and “moving”). This pointed out the necessity of discriminating between **verbal and non-verbal** events to understand how they are temporally related. Our observations suggest that when recognizing an entity semantic type (e.g., sign or symptoms, diseases, and procedures); it is imperative to consider the action associated to it. Therefore, procedures such as colonoscopy, biopsy, pathology, and surgery have to be *performed* (a dynamic verb with a natural end point—an *accomplishment*). Diseases such as adenocarcinoma and appendicitis are *present*, they exist,

Category	C-Start	C-End	NC-Start	NC-End
Activity	+			+
Accomplishment	+	+		
Achievement	+*	+		
State			+	+

** Start and end are so close to each other that this category considers no duration*

Activity	----->	
Accomplishment	-----X	
Achievement		X
State	----->	

Fig. 3 Vendler’s four-way classification. Arrows represent an indefinite time interval, solid lines indicate a homogeneous duration, and dashed lines indicate a dynamic duration. An X is used to represent a situation’s natural end point. Abbreviations: C – Clear and NC – Not Clear.

Fig. 4 Allen and Vendler’s interval representation of OVERLAP and CONTAINS relations. A–/B– and A+/B+ represent the start and end of an event, respectively. Filled-dots represent clear start points while an empty-dot represent a not-clear start point.

Table 5 Sample of the analyzed misclassified sentences by our system

True relation	Predicted relation	Sentence
$Overlap(e_1, e_2)$	$Contains(e_1, e_2)$	1. Tumor <i>invades</i> into the muscularis propria.
$Overlap(e_1, e_2)$	$Contains(e_1, e_2)$	2. Recurrent rectal adenocarcinoma , previously <i>resected</i> node-negative
$Overlap(e_1, e_2)$	$Contains(e_1, e_2)$	3. Mr. Benefield is a pleasant 81-year-old male with resected colon cancer <i>seeking</i> treatment recommendations.
$Overlap(e_1, e_2)$	$Contains(e_1, e_2)$	4. Her chemotherapy was complicated by <i>angina</i> from the 5-FU which was treated with nitroglycerine, and her cardiac evaluation was negative.
$Overlap(e_1, e_2)$	$Contains(e_1, e_2)$	5. This morning, while at dialysis , she had <i>nausea</i> , fevers, and chills.
$Overlap(e_1, e_2)$	$Contains(e_1, e_2)$	6. Exploratory surgery with <i>appendicitis</i> many years ago.
$Overlap(e_1, e_2)$	$Contains(e_1, e_2)$	7. She was seen by a cardiologist in Idyllwild back in April when she was <i>hospitalized</i> and had an adenosine sestamibi scan after that hospitalization, but if surgery is contemplated I would wish her to be seen by cardiology.
$Overlap(e_2, e_1)$	$Contains(e_2, e_1)$	8. Does have some constipation with her iron supplementations but denies nausea, vomiting, abdominal distention, or worsening constipation, as she does have bowel movements <i>once every several days</i> .
$Overlap(e_2, e_1)$	$Contains(e_2, e_1)$	9. She is still moving her bowels <i>multiple times a day</i> .
$Overlap(e_2, e_1)$	$Contains(e_2, e_1)$	10. The patient smokes cigars <i>about once-a-month</i> .

e_1 and e_2 are shown in **bold** and *italics*, respectively.

and consequently, they fall into the *state* category. This is also the case for signs or symptoms like nausea, fever, or discomfort. Following this line of reasoning, it is easier to differentiate an OVERLAP relation from CONTAINS in sentence 5 because we understand that nausea was present during the performance of the dialysis, but there is no enough information as regards to whether the nausea is still present or not. In other words, its end point is unclear. In the case of TIMEX3-EVENT pairs like those in sentences 8 to 10 in Table 5, the nature of the OVERLAP relation between the entities is due to the ambiguity of the time expressions combined with actions that we perceive as ongoing. For example, in sentence 9, the action of moving is an *activity* that is done indeterminably throughout the day as *multiple times a day* imply. On the other hand, in sentence 7, there is a time expression with a definite time interval overlapping the patient's state of being hospitalized. Styler IV et al. (2014) point out that several entities and other non-events are often interpreted in terms of their associated **eventive properties**. However, their discussion differs from ours in that they focus on how these properties define entities such

as medications or disorders as an EVENT, rather than how the implicit interpretation of their eventuality (*taking a medication* or *having a disorder*) is necessary to relate two entities from a temporal perspective. They also introduce the “**contextual aspect**,” which is one of EVENT attributes, but their definition does not relate to the one used in linguistics. The contextual aspect attribute allows one of the three values, N/A, NOVEL, and INTERMITTENT, but as explained in the THYME guidelines, the N/A value simply represents an EVENT as neither NOVEL (i.e., new on the patient’s timeline) nor INTERMITTENT (i.e., when there may be a series of smaller events within a single EVENT). The INTERMITTENT value can be useful in identifying an *activity* or an *accomplishment*, but as shown in Figure 5, just a small portion of EVENTS were annotated with a value different from the default one. Moreover, **aspect is a property of verbs**, and our analysis insinuates that it is more common to find nouns as events.

The temporally locating of two events on a timeline requires a high level of reasoning that even humans can turn into a complicated task. All the aforementioned inferences for differentiating between two of the most frequent and most similar TLINK types (CONTAINS and OVERLAP) were done by heavily relying on the **internal constituency** of an event. Leveraging on aspectual type for temporal relation extraction is a promising approach that was explored by Costa and Branco (2012) on TempEval data, and our analysis implies that clinical data can also profit from it. However, this approach is limited since aspect is a property of verbs.

So far, we have been able to identify a **high similarity** of CONTAINS and OVERLAP relations as one of the reasons why these two types of TLINK are easily confused by our system, which did not pose much difficulties in identifying other TLINK types with a similar number of instances.

Fig. 5 CONTEXTUAL ASPECT attribute values by set.

This differs from what Styler IV et al. (2014) report for the annotator disagreement, which they say comes from different opinions about whether any two EVENTS require an explicit TLINK between them or an inferred one, *rather than what type of TLINK it would be* (e.g., BEFORE vs. CONTAINS). Our observations suggest that the main problem is not the amount of data available, but rather how temporal properties are encoded in language. The next section elaborates this point.

4.3 Temporality of Nominal Events

To deepen our understanding on the complexity of the temporal relation extraction task, we divided all OVERLAP and CONTAINS false negatives into the four possible pair types: EVENT-EVENT, TIMEX3-TIMEX3, EVENT-TIMEX3, and TIMEX3-EVENT. As shown in Table 6 (left), a significant amount of OVERLAP and CONTAINS links were EVENT-EVENT relations. Therefore, we looked further into this type of pairs, discriminating between verb (V) and non-verb (NV) events. Table 6 (right) shows the results in more detail.

As mentioned by Pustejovsky and Stubbs (2011) and further discussed in Styler IV et al. (2014), EVENT-EVENT pairings are a complex and vital component, particularly in clinical narratives, where doctors rely on shared domain knowledge and it is essential to read “between the lines.” The distribution of verb/non-verb entities in Table 6 (right) indicates that most EVENT-EVENT misclassified pairings were either of NV-NV type or included a NV entity. This finding is of prime relevance to temporal reasoning since temporality is naturally encoded in verbs, expressing actions or events, while nouns are usually the person or thing doing or receiving that action (i.e., subject or object). Without a verb, the semantics of nouns hardly give a notion of time. Consider the following sentences:

- (1) **Tumor** *invades* into the muscularis propria.
- (2) Resected cecal **adenocarcinoma** with *resection* of liver metastasis.

Table 6 Distribution of misclassified CONTAINS and OVERLAP pairs by type of TLINK (left) and EVENT type (right)

Dev set: TLINK pairs					Dev set: Event-Event pairs				
TLINK	E-E	T-T	E-T	T-E	TLINK	V-V	V-NV	NV-V	NV-NV
CONTAINS	149	6	2	37	CONTAINS	5	42	24	78
OVERLAP	251	0	25	46	OVERLAP	3	54	24	170
Total	400	6	27	83	Total	8	96	48	248

Abbreviations: E–EVENT, T–TIMEX3, V–Verb and NV–Non-Verb

In sentence 1, the NV entity “tumor” is annotated to overlap with the V entity “invades.” Similarly, in sentence 2 “adenocarcinoma” is annotated to overlap with “resection.” To understand how the time interval of these EVENT entities overlap, we inevitably look for an associated action to picture their duration, but the sole definition of *tumor* or *adenocarcinoma* does not provide us with that information. Clearly, it is not until we attribute these two nouns the property of *being present* (i.e., to exist) that we can think about a *state*, which has no inherent endpoint. This forced reasoning is not straightforward, even for humans. This depicts the fact that it could be even harder for computers to process.

Considering that verbs can be nominalized, we looked for nominalizations in our system’s misclassified sentences. We observed some NV entities such as “consultation,” “diagnosis,” “discharge,” “examination,” and “resection,” which derive from the verbs “to consult,” “to diagnose,” “to discharge,” “to examine,” and “to resect.” However, it was more common to find NV entities like “cancer,” “diabetes,” “history,” “anesthesia,” and “dialysis,” just to name a few. These entities are a good example of non-events being interpreted in terms of their (implicit) associated action. The THYME project defines an EVENT as *anything* relevant to the clinical timeline. This interpretation is broader than the one originally defined by Pustejovsky, Ingria, Sauri, Castano, Littman, Gaizauskas, Setzer, Katz, and Mani (2005), where the term EVENT considers anything that *happens or occurs*, and is generally expressed by means of tensed or untensed verbs, nominalizations, adjectives, predicative clauses, or prepositional phrases. Consequently, the THYME project definition allows non-events, such as medications or disorders, to be annotated as EVENTS. While Styler IV et al. (2014) mentioned this, they did not show the frequency of nouns.

As we saw in previous examples, the time intervals of NV entities are more difficult to conceptualize, while V entities, such as “removed” or “improving,” have their time properties morphologically encoded. Therefore, regardless of the low number of V-V relations, temporal information from verb predicates usually have more explicit hints; NV entities are more challenging and require more careful examination.

In Section 4.2.2, we pointed out the high similarity of OVERLAP and CONTAINS as one of the challenges of the temporal relation extraction task. Here we conclude that the high frequency of NV entities and the complexity of noun-noun relations is likely to be another reason why our system and previous works lag behind human performance. Not even the model of Miwa and Bansal (2016), which was designed to extract noun-noun relations, was able to handle the TLINKS in Clinical TempEval. As was noted earlier, the semantics of nouns are not enough to give the notion of an EVENT duration. This directly affects our system’s performance.

We already introduced Vendler’s aspectual classification and discussed how it helps to sepa-

rate two extremely similar TLINKS. Unfortunately, this is not compatible with nominal predicates. In order to be able to use Vendler’s topology in the clinical domain, the currently implicit associated actions to nominal EVENTS would have to be manually made explicit. Assuming that once identified, the verbs were classified as an *activity*, *accomplishment*, *achievement*, or *state*, this information can be used as a feature vector by our model. Alternatively, verb/non-verb entities distinction of EVENTS is the first step that can alleviate the incompatibility of aspect with nominals, and positively influence the temporal relation extraction task.

4.4 Precision and Recall Imbalance

All the experiments introduced in Section 2.2 outperformed the best Clinical TempEval 2016 system and the state-of-the-art model as shown in Table 3 and Table 4. However, in all the settings, our model showed high performance, save for a recall lower than all of the systems presented in Table 3. Previously, we attributed this to the high number of negative instances in our dataset. As seen in Table 4, filtering some of the negative instances resulted in a slight increase in recall, but we were still unable to reach a better balance with precision. Moreover, our best recall score was still below the one achieved by Lee et al. (2016) and Lin et al. (2016). Consequently, we present an additional set of experiments to complement the multi-classification results in Table 4; i.e., training the model without the NONE class.⁶ These results are shown in Table 7.

We can observe that the overall performance of our model remains the same. The classification of CONTAINS remains the highest, followed by BEFORE. The performance of the ENDS-ON and OVERLAP varies between the third and fourth best: ENDS-ON is the third best under the

Table 7 Results of our multi-class classification experiments on the THYME test set

TLINK	Multi-class classification											
	<i>Wikipedia word emb</i>			<i>Wikipedia without None</i>			<i>PubMed word emb</i>			<i>PubMed without None</i>		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BEFORE	0.696	0.183	0.289	0.573	0.461	0.511	0.704	0.196	0.306	0.567	0.466	0.511
BEGINS-ON	0.628	0.082	0.145	0.462	0.284	0.351	0.620	0.110	0.186	0.437	0.302	0.357
CONTAINS	0.907	0.468	0.617	0.810	0.647	0.719	0.904	0.471	0.619	0.814	0.641	0.717
ENDS-ON	0.525	0.093	0.157	0.551	0.325	0.408	0.656	0.122	0.204	0.462	0.306	0.368
OVERLAP	0.494	0.121	0.195	0.415	0.352	0.381	0.526	0.124	0.201	0.404	0.346	0.373
Macro-F1	0.281			0.474			0.303			0.465		

Our results come from five different random seeds. *Without None* refers to training without the None class.

⁶ We did not remove the NONE class from the test set.

Wikipedia without None and *PubMed word embeddings*, while OVERLAP is under the *Wikipedia word embeddings* and *PubMed without None*. BEGINS-ON showed the lowest performance. Removing the NONE class resulted in a better balance between precision and recall, increasing the latter for all classes. Consequently, the F1 score increased as well. Despite this change, the OVERLAP still showed the second lowest performance.

4.4.1 Class oversampling

At the beginning of these experiments, we assumed that the dataset size was not one of the underlying reasons of the low performance witnessed in previous works. Our error analysis indicated that low performance can be attributed to two reasons: the high similarity of OVERLAP with CONTAINS and the temporality of nominal events. Since we are using a neural model, it is natural to think that the more instances we have, the better the classification. We explored this assumption using four new experimental settings:

- (I) **Oversampling OVERLAP:** Given that OVERLAP has one of the lowest performance, even though it is the second most frequent class in the dataset, we balance OVERLAP with CONTAINS by random oversampling instances in the train set (ratio 1:1).
- (II) **Oversampling OVERLAP (without None):** Same setting as (I). The NONE class is removed for training but we maintain it for purposes of testing.
- (III) **Oversampling all the minority classes:** For better comparison, we balance all the minority classes (BEFORE, BEGINS-ON, ENDS-ON, and OVERLAP) with CONTAINS by random oversampling instances in the train set (ratio 1:1).
- (IV) **Oversampling all the minority classes (without None):** Same setting as (III). The NONE class is removed for training but we maintain it for purposes of testing.

The results are shown using default Wikipedia word embeddings and PubMed word embeddings in Table 8. Once again, we observe that the overall performance of our model remained the same. In the best case scenario, OVERLAP has the third best performance. However, it continues to show similar performance with the remaining minority classes. Similar to our first set of experiments, training with the NONE class results in high precision but low recall. In line with the results of Table 7, by removing the NONE class from the training set we get a better balance. Both oversampling techniques resulted in close Macro-F1 scores whenever we removed the NONE class from training and changed the default Wikipedia word embeddings for PubMed word embeddings (0.449 vs. 0.449 and 0.468 vs. 0.465, respectively). A McNemar’s test (McNemar 1947) on these results yields a p-value of 0.077 for *Wikipedia without None* and *PubMed without None* oversampling OVERLAP and a p-value of 0.466 when oversampling all minority classes.

Table 8 Results of our multi-class classification experiments on the THYME test set

TLINK	Multi-class classification oversampling OVERLAP											
	<i>Wikipedia word emb</i>			<i>Wikipedia without None</i>			<i>PubMed word emb</i>			<i>PubMed without None</i>		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BEFORE	0.686	0.165	0.266	0.592	0.423	0.494	0.679	0.155	0.252	0.571	0.447	0.501
BEGINS-ON	0.593	0.090	0.157	0.478	0.253	0.331	0.727	0.082	0.148	0.416	0.242	0.306
CONTAINS	0.900	0.461	0.610	0.810	0.631	0.709	0.909	0.448	0.601	0.814	0.618	0.703
ENDS-ON	0.583	0.046	0.086	0.500	0.265	0.346	0.684	0.086	0.153	0.490	0.318	0.386
OVERLAP	0.365	0.164	0.227	0.355	0.373	0.363	0.373	0.188	0.250	0.344	0.360	0.351
Macro-F1	0.269			0.449			0.281			0.449		

TLINK	Multi-class classification oversampling all minority classes											
	<i>Wikipedia word emb</i>			<i>Wikipedia without None</i>			<i>PubMed word emb</i>			<i>PubMed without None</i>		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BEFORE	0.629	0.287	0.394	0.549	0.446	0.492	0.622	0.287	0.393	0.532	0.469	0.499
BEGINS-ON	0.466	0.209	0.288	0.375	0.340	0.357	0.392	0.201	0.266	0.325	0.312	0.318
CONTAINS	0.912	0.449	0.602	0.827	0.614	0.705	0.923	0.441	0.597	0.836	0.603	0.701
ENDS-ON	0.448	0.371	0.406	0.424	0.424	0.424	0.456	0.344	0.392	0.415	0.450	0.432
OVERLAP	0.409	0.165	0.235	0.368	0.354	0.361	0.410	0.164	0.234	0.379	0.367	0.373
Macro-F1	0.385			0.468			0.376			0.465		

The first results are results obtained when oversampling only OVERLAP, and the subsequent results are obtained when oversampling BEFORE, BEGINS-ON, ENDS-ON, and OVERLAP.

This is not significant at the 0.05 alpha level. The reason might be the same as why in-domain embeddings had a limited improvement in Section 2.2 multi-class classification experiments (see Appendix B.1).

5 Conclusions

Clinical language processing represents a special challenge to NLP systems. The structure of clinical texts range from telegraphic constructions to long utterances describing a patient’s condition or a suggested diagnosis. The high use of domain knowledge to infer temporal relations between events does not make this task any easier. A doctor naturally interprets adenocarcinoma (a type of cancer) as an abnormal, uncontrolled and *progressive* growth of tissue, which temporally speaking is and should be thought as an ongoing process unless explicitly qualified (“*We resected the adenocarcinoma, and since margins were clear, we can say it is gone*”). This is a non-trivial task for a computer even when relying on context information.

There have been several attempts on tackling temporal relation extraction from clinical text, mostly led by the Clinical TempEval challenges. However, the results are still far from human

performance and there is little information about the underlying reasons. This encouraged our work to adapt a state-of-the-art system and do a detailed error analysis, which pointed out that one of the major challenges is how to handle the eventive properties of nominals—the predominant type of events on the most frequent type of pairs (EVENT-EVENT).

Existing knowledge bases, such as the Unified Medical Language System’s (UMLS) Metathesaurus help to classify entities into semantic types like *Therapeutic or Preventive procedure*, *Sign or Symptom* or *Disease or Syndrome*. However, the associated events and actions cannot be found in this or any other knowledge base. Therefore, we hypothesize that a resource containing aspectual information of the actions associated to common nominals, such as procedures or diseases, can further improve temporal relation extraction in the clinical domain. Since this will require manual annotation effort from annotators with linguistic and clinical knowledge, we first plan to analyze further EVENT-EVENT relations by differentiating events as verbal and non-verbal events.

Acknowledgement

This work was supported by JST CREST Grant Number JPMJCR1513, Japan. We would like to thank the anonymous reviewers for their insightful comments and suggestions.

Reference

- Allen, J. F. (1990). “Readings in Qualitative Reasoning About Physical Systems.”. Chap. Maintaining Knowledge About Temporal Intervals, pp. 361–372. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- Andersen, R. W. (1990). *Unpublished Lecture in the Seminar on the Acquisition of Tense and Aspect*. Applied Linguistics, University of California.
- Bethard, S., Derczynski, L., Savova, G., Pustejovsky, J., and Verhagen, M. (2015). “SemEval-2015 Task 6: Clinical TempEval.” In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 806–814, Denver, Colorado. Association for Computational Linguistics.
- Bethard, S., Savova, G., Chen, W.-T., Derczynski, L., Pustejovsky, J., and Verhagen, M. (2016). “SemEval-2016 Task 12: Clinical TempEval.” In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 1052–1062, San Diego, California. Association for Computational Linguistics.

- Bethard, S., Savova, G., Palmer, M., and Pustejovsky, J. (2017). “SemEval-2017 Task 12: Clinical TempEval.” In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 565–572, Vancouver, Canada. Association for Computational Linguistics.
- Bybee, J. L., Perkins, R. D., and Pagliuca, W. (1994). *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*, Vol. 196. University of Chicago Press Chicago.
- Chikka, V. R. (2016). “CDE-IIITH at SemEval-2016 Task 12: Extraction of Temporal Information from Clinical documents using Machine Learning techniques.” In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 1237–1240, San Diego, California. Association for Computational Linguistics.
- Costa, F. and Branco, A. (2012). “Aspectual Type and Temporal Relation Classification.” In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 266–275, Avignon, France. Association for Computational Linguistics.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). “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 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Dligach, D., Miller, T., Lin, C., Bethard, S., and Savova, G. (2017). “Neural Temporal Relation Extraction.” In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 746–751, Valencia, Spain. Association for Computational Linguistics.
- Fries, J. (2016). “Brundlefly at SemEval-2016 Task 12: Recurrent Neural Networks vs. Joint Inference for Clinical Temporal Information Extraction.” In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 1274–1279, San Diego, California. Association for Computational Linguistics.
- Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., and Szpakowicz, S. (2010). “SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals.” In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 33–38, Uppsala, Sweden. Association for Computational Linguistics.
- Jung, H., Allen, J., Blaylock, N., de Beaumont, W., Galescu, L., and Swift, M. (2011). “Building Timelines from Narrative Clinical Records: Initial Results Based-on Deep Natural Language

- Understanding.” In *Proceedings of BioNLP 2011 Workshop*, pp. 146–154, Portland, Oregon, USA. Association for Computational Linguistics.
- Kingma, D. P. and Ba, J. (2014). “Adam: A Method for Stochastic Optimization.” *arXiv preprint arXiv:1412.6980*.
- Klein, W. (2013). *Time in Language*. Routledge.
- Lee, H.-J., Xu, H., Wang, J., Zhang, Y., Moon, S., Xu, J., and Wu, Y. (2016). “UTHealth at SemEval-2016 Task 12: an End-to-End System for Temporal Information Extraction from Clinical Notes.” In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 1292–1297, San Diego, California. Association for Computational Linguistics.
- Leeuwenberg, A. and Moens, M.-F. (2016). “KULeuven-LIIR at SemEval 2016 Task 12: Detecting Narrative Containment in Clinical Records.” In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 1280–1285, San Diego, California. Association for Computational Linguistics.
- Leeuwenberg, A. and Moens, M.-F. (2017). “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*, pp. 1150–1158, Valencia, Spain. Association for Computational Linguistics.
- Li, P. and Shirai, Y. (2000). *The Acquisition of Lexical and Grammatical Aspect*. Studies on Language Acquisition. Mouton de Gruyter, Berlin, New York.
- Li, P. and Huang, H. (2016). “UTA DLNLP at SemEval-2016 Task 12: Deep Learning Based Natural Language Processing System for Clinical Information Identification from Clinical Notes and Pathology Reports.” In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 1268–1273, San Diego, California. Association for Computational Linguistics.
- Lin, C., Miller, T., Dligach, D., Bethard, S., and Savova, G. (2016). “Improving Temporal Relation Extraction with Training Instance Augmentation.” In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pp. 108–113, Berlin, Germany. Association for Computational Linguistics.
- McNemar, Q. (1947). “Note on the Sampling Error of the Difference between Correlated Proportions or Percentages.” *Psychometrika*, **12** (2), pp. 153–157.
- Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013). “Efficient Estimation of Word Representations in Vector Space.”.
- Miwa, M. and Bansal, M. (2016). “End-to-End Relation Extraction using LSTMs on Sequences

- and Tree Structures.” In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1105–1116, Berlin, Germany. Association for Computational Linguistics.
- Pustejovsky, J., Ingria, B., Sauri, R., Castano, J., Littman, J., Gaizauskas, R., Setzer, A., Katz, G., and Mani, I. (2005). “The Specification Language TimeML.” *The Language of Time: A Reader*, pp. 545–557.
- Pustejovsky, J. and Stubbs, A. (2011). “Increasing Informativeness in Temporal Annotation.” In *Proceedings of the 5th Linguistic Annotation Workshop*, pp. 152–160, Portland, Oregon, USA. Association for Computational Linguistics.
- Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G., Lehman, L.-W., Moody, G., Heldt, T., Kyaw, T. H., Moody, B., and Mark, R. G. (2011). “Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A Public-Access Intensive Care Unit Database.” *Critical Care Medicine*, **39** (5), pp. 952–960.
- Smith, C. S. (1983). “A Theory of Aspectual Choice.” *Language*, pp. 479–501.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). “Dropout: A Simple Way to Prevent Neural Networks from Overfitting.” *Journal of Machine Learning Research*, **15** (56), pp. 1929–1958.
- Styler IV, W. F., Bethard, S., Finan, S., Palmer, M., Pradhan, S., de Groen, P. C., Erickson, B., Miller, T., Lin, C., Savova, G., and Pustejovsky, J. (2014). “Temporal Annotation in the Clinical Domain.” *Transactions of the Association for Computational Linguistics*, **2**, pp. 143–154.
- UzZaman, N., Llorens, H., and Allen, J. (2012). “Evaluating Temporal Information Understanding with Temporal Question Answering.” In *2012 IEEE Sixth International Conference on Semantic Computing*, pp. 79–82.
- UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., and Pustejovsky, J. (2013). “SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations.” In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pp. 1–9, Atlanta, Georgia, USA. Association for Computational Linguistics.
- Vendler, Z. (1957). “Verbs and Times.” *The Philosophical Review*, **66** (2), pp. 143–160.
- Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., and Pustejovsky, J. (2007). “SemEval-2007 Task 15: TempEval Temporal Relation Identification.” In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 75–80, Prague, Czech Republic. Association for Computational Linguistics.

- Verhagen, M., Saurí, R., Caselli, T., and Pustejovsky, J. (2010). “SemEval-2010 Task 13: TempEval-2.” In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 57–62, Uppsala, Sweden. Association for Computational Linguistics.
- Werbos, P. J. (1990). “Backpropagation through Time: What It Does and How to Do It.” *Proceedings of the IEEE*, **78** (10), pp. 1550–1560.
- Xu, K., Feng, Y., Huang, S., and Zhao, D. (2015). “Semantic Relation Classification via Convolutional Neural Networks with Simple Negative Sampling.” In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 536–540, Lisbon, Portugal. Association for Computational Linguistics.

Appendix

A Adaptation

A.1 Sentence-level Annotations

We used the Clinical Language Annotation, Modeling and Processing (CLAMP) toolkit⁷ for tokenization and sentence boundary detection. We then matched all entities’ spans from the gold standard with the sentence offsets on the CLAMP output to identify those within the same sentence. Therefore, we created new annotations containing a pair of words, their offsets in the sentence, the temporal relation between them marked on the gold standard, and the directionality of the arguments. Example 3 shows an annotation of the TLINK—CONTAINS(*lifelong*, *nonsmoker*) in the sentence “*He is a lifelong nonsmoker.*” Note that no entity type (i.e., EVENT or TIMEX3) or any of its associated attributes are included. Our code for generating these sentence-level annotations is available at <https://github.com/cl-tohoku/thyme-sentence-annotations>.⁸

	T1	Term 8 16	lifelong
(3)	T2	Term 17 26	nonsmoker
	R1	ContainsSource-ContainsTarget	Arg1:T1 Arg2:T2

The THYME corpus does not identify instances where two entities have none of the TLINK relations. Hence, we define a NONE label and apply it as follows: since any two EVENT/TIMEX3

⁷ <http://clinicalnlp-tool.com/index.php>

⁸ Note that it will be necessary to request access to the THYME corpus. Please follow the instructions on the THYME wiki: https://clear.colorado.edu/TemporalWiki/index.php/Main_Page#Getting_access_to_the_THYME_corpus_and_gold_standard_annotations

can be a candidate pair, we take all entities in a sentence to generate all pair permutations as candidates. Pairs that do not have any temporal relation are then labeled as NONE. Therefore, for entities e_1 , e_2 , and e_3 in a sentence where e_1 CONTAINS e_2 , pair (e_1, e_2) is considered as a positive instance while the resulting candidate pairs from our procedure (e_1, e_3) , (e_2, e_1) , (e_2, e_3) , (e_3, e_1) , and (e_3, e_2) are considered as negative instances. Due to the large number of negative instances produced (1:3 ratio of positive to negative examples), for a sentence with entities e_4 , e_5 , and e_6 with a TLINK: e_5 BEFORE e_6 —(or any TLINK but CONTAINS)—no negative instances were generated.

A.2 Implementation and Training

This study used Miwa and Bansal’s (2016) implementation, available at <https://github.com/tticoin/LSTM-ER>. It followed their training settings, updating the model parameters (including weights, biases, and embeddings) by backpropagation through time (Werbos 1990) and Adam (Kingma and Ba 2014) with gradient clipping, parameter averaging, and L2-regularization. Dropout (Srivastava, Hinton, Krizhevsky, Sutskever, and Salakhutdinov 2014) was applied to the embedding layer and to the final hidden layers for relation classification. The hyper-parameters used were their default hyper-parameters for the SemEval-2010 Task 8: initial learning rate (1e-6), regularization parameter (1e-6), input dropout probability (0.5), output dropout probability (0.3), gradient clipping size (1), and number of epochs (63).

B Multi-Classification Performance

B.1 In-Domain Word Embeddings

Once we verified the adopted model gave competitive results on the narrative container identification task, we focused on increasing the system’s recall. We, therefore changed the default word representations trained on Wikipedia for in-domain word embeddings. Word representation depends on the words in context, and since the clinical domain is a specific field with different vocabulary from those used in the general domain, we expected the model to benefit from a resource like PubMed. However, our results suggest that this does not have a significant impact on most TLINKS. Only BEGINS-ON and ENDS-ON recall considerably improved. This limited improvement can be attributed to the data size. The subset of PubMed abstracts used to train our in-domain word embeddings is smaller than the Wikipedia dump on which Miwa and Bansal’s (2016) default word embeddings were trained. A possible method of improving word representation in a temporal-aware context is to rely on transfer learning. For this purpose, pre-training BERT

(Devlin, Chang, Lee, and Toutanova 2019) on general domain temporal data and fine-tuning on the Clinical TempEval task could lead to interesting results.

B.2 Down-Sampling Negative Examples

We still witnessed an imbalance between precision and recall despite the fact that we increased recall by using in-domain word embeddings. Moreover, our results are still below UHealth’s recall score (highest on CONTAINS identification task). By filtering EVENT pairs as described in Section 2.2 experimental setting (IV), the NONE class was reduced by 10%. This further improved the recall for most TLINKS except for ENDS-ON. A McNemar’s test on the results of *PubMed word emb* and *PubMed word emb + FNE* yields a p-value of 0.006, which is significant at the 0.05 level. This means that by filtering negative examples, the model’s proportion of errors decreases with respect to *PubMed word emb*.

Diana Galvan-Sosa: Received the bachelor’s degree in computer science and technology from the Monterrey Institute of Technology and Higher Education, Mexico, in 2010, and the M.Sc degree in information science from Tohoku University, Japan, in 2018, where she is currently pursuing the Ph.D. degree in information science. Her research interests include natural language processing, especially knowledge acquisition for machine reading comprehension.

Koji Matsuda: Received his bachelor’s degree in engineering from Toyohashi Institute of Technology in 2006, and his M.E. degree in engineering from the Tokyo Institute of technology in 2012. He is currently a Ph.D. candidate at the Graduate School of Information Sciences, Tohoku University. He also works as Technical Staff in RIKEN since 2018. His research interests include natural language processing, especially automatic knowledge base construction and relation extraction.

Naoaki Okazaki: Professor at Department of Computer Science, School of Computing, Tokyo Institute of Technology. He obtained his Ph.D. from Graduate School of Information Science and Technology, University of Tokyo in 2007. He worked as a post-doctoral researcher in University of Tokyo (in 2007–2011) and as an associate professor in Tohoku University (in 2011–2017). He is also a visiting research scholar of the Artificial Intelligence Research Center (AIRC), AIST. His research interests include natural language processing, text mining,

and machine learning.

Inui Kentaro: Professor of the Graduate School of Information Sciences at Tohoku University, where he is head of the Natural Language Processing Lab. He also leads the Natural Language Understanding Team at the RIKEN Center for the Advanced Intelligence Project. His research areas include natural language processing and artificial intelligence. He currently serves as Vice-chairperson of Association for Natural Language Processing, Member of Science Council of Japan, and Director of NPO FactCheck Initiative Japan.

(Received November 1, 2019)

(Revised February 7, 2020)

(Accepted March 13, 2020)