Investigating the Challenges of Temporal Relation Extraction from Clinical Text

Diana Galvan [1] **Naoaki Okazaki** [2] **Koji Matsuda** [1] **Kentaro Inui** [1,3]

[1] Tohoku University [2] Tokyo Institute of Technology

[3] RIKEN Center for Advanced Intelligence Project

{dianags,matsuda,inui}@ecei.tohoku.ac.jp

okazaki@dc.titech.ac.jp

Abstract

Temporal reasoning remains as an unsolved task for Natural Language Processing (NLP), particularly demonstrated in the clinical domain. The complexity of temporal representation in language is evident as results of the 2016 Clinical TempEval challenge indicate: the current state-of-the-art systems perform well in solving mention-identification tasks of event and time expressions but poorly in temporal relation extraction, showing a gap of around 0.25 point below human performance. We explore to adapt the tree-based LSTM-RNN model proposed by Miwa and Bansal (2016) to temporal relation extraction from clinical text, obtaining a five point improvement over the best 2016 Clinical TempEval system and two points over the state-of-the-art. We deliver a deep analysis of the results and discuss the next step towards human-like temporal reasoning.

1 Introduction

Temporal Information Extraction (TIE) is an active research area in NLP, where the ultimate goal is to be able to represent the development of a story over time. TIE is a key to text processing tasks including Question Answering and Text Summarization and follows the traditional pipeline of named entity recognition (NER) and relation extraction separately. Research on this area has been led by TempEval shared tasks (Verhagen et al., 2007, 2010; UzZaman et al., 2013) but in recent years, the target domain has been shifted to the clinical domain. The resulting Clinical TempEval challenges (Bethard et al., 2015, 2016, 2017) introduced the adoption of narrative containers to their annotation schema, based on the widely used TIE annotation standard ISO-TimeML (Pustejovsky et al., 2010). Narrative containers were 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 series of events may fall. They help visualize the temporal relations within a text and facilitate the identification of other temporal relation types. Until now, the only corpus annotated with narrative containers is limited to clinical texts.

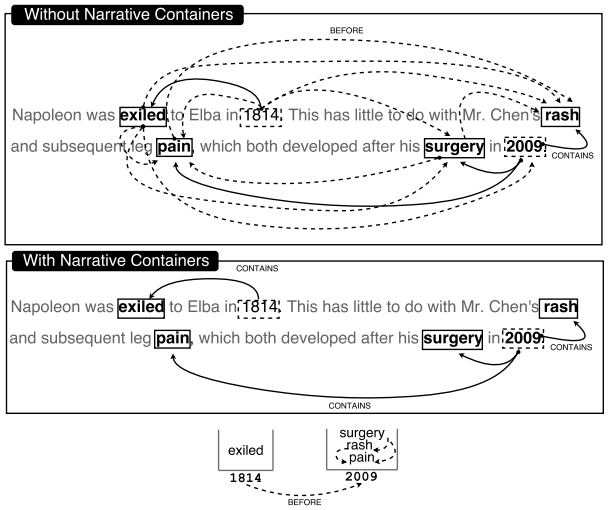

Figure 1: Example temporal relation annotation with and without using narrative containers.

Results of the systems participating in Clinical TempEval suggest that they perform well on time-entity identification tasks. Nevertheless, temporal relation extraction has shown to be the most difficult. UTHealth (Lee et al., 2016), the best ranked system in 2016 Clinical TempEval, showed a significant gap of 0.25 when compared to human performance even with gold-standard entity annotations. Recent work by Lin et al. (2016) and Leeuwenberg and Moens (2017) improved UTHealth's results further but the gap with respect to humans is still around 0.21. Regardless of the

Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis (LOUHI 2018), pages 55–64
Brussels, Belgium, October 31, 2018.

increase in annotation agreement of temporal relations by relying on narrative containers, there is a consensus within the research community regarding TIE difficulty. Still, the reasons of the uneven results between entity and temporal relation predictions remain unclear.

We attribute the complexity of temporal representation in natural language as the main cause of the low performance on temporal relation tasks. Tense and aspect are the two grammatical means to express the notion of time in English but little has been discussed about the latter on clinical text. Furthermore, the focus of previous work on temporal relation extraction is set on narrative containers, which have proved to be useful to locate and relate two events on a timeline. Identification of other temporal relation types has been less frequently tackled. We believe is key to look at the whole set of temporal types to achieve the ultimate goal of developing systems that automatically create a timeline of a patient's health care.

In this paper, we describe the process followed to adapt the neural model proposed by Miwa and Bansal (2016) on TIE, which has already shown competitive results on semantic relation extraction. In our pursuit of understanding the nature of the challenges that characterize the processing of temporal relations, we continue with an error analysis of our system's overall performance and not only on the identification of narrative containers. Our final goal is to shed some light on the difficulties of temporal relation extraction and the necessary efforts to improve further current state-of-the-art systems performance with that of humans on completing the same task.

2 Related Work

Due to the recent shift of TIE to the clinical domain, most related work has been done by Clinical TempEval participating systems. This challenge uses a corpus annotated with five different temporal relation (TLINK) types between events and times ("TIMEX3" in this schema): BEFORE, BEGINS-ON, CONTAINS, ENDS-ON and OVERLAP. However, this challenge only evaluates the identification of a narrative container, marked with the CONTAINS type.

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

Recent work by Lin et al. (2016), Dligach et al. (2017) and Leeuwenberg and Moens (2017) followed the settings of 2016 Clinical TempEval challenge but they did not participate in the competition. Out of these, our results are only directly comparable to those of Lin et al. (2016) and Leeuwenberg and Moens (2017) since the work of Dligach et al. (2017) was not evaluated using the Clinical TempEval official scorer.

Even though Leeuwenberg and Moens (2017) established a new state-of-the-art in temporal relation extraction, their result is still below human performance. Moreover, none of the aforementioned works provides a detailed discussion of *why* is current performance so low and *how* can we improve further the results on temporal relation extraction, except from Leeuwenberg and Moens (2016), which in their first attempt on tackling this task on 2016 Clinical TempEval identified false negatives as their major problem.

Our contribution is a deep error analysis taking into account the performance of our model on predicting all TLINK types. As a result, we were able to identify important clues on temporal relation extraction and based on these findings, we discuss the next step towards human-like temporal reasoning performance.

3 Method

We adapted the tree-based bidirectional LSTM-RNN end-to-end neural model of Miwa and Bansal (2016) to intra-sentential temporal relation extraction from clinical text. This three-layer model (embedding, sequence and dependency layers) jointly identifies entities and relations between them. For relation classification, the model heavily relies on the dependency structure around the target word pair and the output of the sequence

TLINK	Train	Test
CONTAINS	8653	4554
NONE	43643	20465
Total	52296	25019

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

TLINK	Train	Test
BEFORE	1839	982
BEGINS-ON	717	363
CONTAINS	8653	4554
ENDS-ON	334	138
OVERLAP	2388	1186
NONE	43643	20465
Total	57574	27688

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

System	P	R	F1
(Lee et al., 2016)	0.588	0.559	0.573
(Lin et al., 2016)	0.669	0.534	0.594
(Leeuwenberg and Moens, 2017)	-	-	0.608
Our model	0.983	0.462	0.629
Human performance	-	-	0.817

Table 3: Performance of systems and humans on identifying CONTAINS relations.

layer. When tested on nominal relation classification (Hendrickx et al., 2009), it showed competitive results against the state-of-the-art.

We followed the official 2016 Clinical TempEval settings for phase 2 of evaluation, where given the raw text and manual event and time annotations, the task is to identify the temporal relation between a directed pair (e_1, e_2), if any. e_1 and e_2 are entities of either EVENT or TIMEX3 type. For relation classification, Miwa and Bansal (2016) model takes as an input a sentence and a annotation file with a word pair. The output contains the predicted relation type and the directionality of the entities:(e_1, e_2) when e_1 is the source and e_2 the target and (e_2, e_1) otherwise.

4 Experimental settings

4.1 Dataset

Similar to 2016 Clinical TempEval, we used the THYME corpus (Styler IV et al., 2014) for evaluation, 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 identified entities are given a set of attributes depending on their type: *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.

Sentence-level annotations are necessary to meet Miwa and Bansal (2016)'s input requirements. Therefore, we used the Clinical Language Annotation, Modeling and Processing (CLAMP) toolkit[1] for tokenization and sentence boundary detection. We matched all entities spans from the gold standard with the sentence offsets on the CLAMP output to identify those within the same sentence. As a result, the new annotations contain 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 1 shows an example annotation of the TLINK CONTAINS*(lifelong, nonsmoker)* in the sentence *He is a lifelong nonsmoker*.

(1)

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

Since any two EVENT/TIMEX3 can be a candidate pair, we took all entities in a sentence to generate all pair combinations as candidates. Pairs that do not have any temporal relation were labeled as NONE. Due to the large number of negative instances produced by this procedure, it was applied only to CONTAINS. No negative instances were generated for the remaining TLINK types and we did not extend the set of TLINKs to its transitive closure (i.e. A CONTAINS B $\wedge$ B CONTAINS C $\rightarrow$ A CONTAINS C). Table 1 and Table 2 detail the resulting datasets.

[1] http://clinicalnlptool.com/index.php

	Binary classification			Multi-class classification								
	Wikipedia word emb			*Wikipedia word emb*			*PubMed word emb*			*PubMed word emb + FNE*		
TLINK	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BEFORE	-	-	-	0.698	0.185	0.292	0.708	0.198	0.310	0.683	0.202	0.312
BEGINS-ON	-	-	-	0.585	0.062	0.112	0.615	0.103	0.177	0.608	0.116	0.195
CONTAINS	0.983	0.462	**0.629**	0.905	0.472	**0.621**	0.908	0.471	**0.620**	0.889	0.479	**0.623**
ENDS-ON	-	-	-	0.520	0.086	0.148	0.704	0.126	0.213	0.760	0.126	0.216
OVERLAP	-	-	-	0.504	0.134	0.211	0.504	0.134	0.211	0.497	0.140	0.218

Table 4: Results of our four experiments on the THYME test set. FNE refers to filtered negative examples.

4.2 Experiments

We followed the same experimental settings described in Miwa and Bansal (2016). Additional to the model's default Wikipedia word embeddings, we trained word vectors of 200 dimensions using word2vec (Mikolov et al., 2013) on a subset of journal abstracts in Oncology and Gastroenterology from PubMed2014[2]. PubMed data can be easily downloaded without application approval that clinical corpus like MIMIC II (Saeed et al., 2011) require.

We conducted four experiments at the intra-sentential level. The first experiment follows 2016 Clinical TempEval, focusing only on the identification of the CONTAINS type. The remaining experiments include the five annotated TLINKs. Further detail of each experiment is given below:

1. TLINK:CONTAINS binary classification: In order to obtain results comparable to Lee et al. (2016), the best ranked system in 2016 Clinical TempEval, we only considered TLINK:CONTAINS instances. The model chooses between CONTAINS and NONE relations.

2. Multi-class classification: To test the model in a real-world setting, we added to train and test sets the remaining pairs in the gold standard that have any of the other TLINK types. No further negative examples were created for the additional types.

3. Multi-class classification with PubMed word embeddings: In addition to the previous setting (2), we used word embeddings trained on the subset of PubMed instead of the default word vectors trained on Wikipedia.

4. Multi-class classification with PubMed word embeddings and filtered negative examples: In addition to the previous setting (3), we filtered from the dataset NONE pairs that according to the THYME guidelines[3] should never be TLINKed. Thus, we removed a candidate pair whenever e_1 contextual modality value[4] was ACTUAL or HEDGED and the e_2 had HYPOTHETICAL or GENERIC modality, and vice versa.

5 Results

5.1 TLINK:CONTAINS binary classification

Table 3 presents the results of previous approaches compared to human performance. The first row shows the top performance in 2016 Clinical TempEval using binary classification. The second and third rows are the latests results outside the competition. Following the steps of the Clinical TempEval narrative container identification task, we only tried to predict TLINKs of CONTAINS type. In doing so we obtained an F1 score of 0.629, outperforming UTHealth's system. The model shows a high precision but lower recall than UTHealth; this is probably because of NONE relations prevailing in the dataset. By handling the task as binary classification, given a pair of entities we are already assuming there is some kind of temporal relation and the classifier's task is to decide whether it is CONTAINS or not. We performed this experiment in order to have results comparable with those of UTHealth. However, we cannot compare this re-

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

[3] http://savethevowels.org/files/THYMEGuidelines.pdf, Section 6.2.5

[4] Entity attributes introduced in Section 4.1 were not used as features in our model. EVENTs marked with HYPOTHETICAL or GENERIC modality are non-real events. Therefore, they cannot be related to real events marked as ACTUAL or HEDGED.

sult to the state-of-the-art since Leeuwenberg and Moens (2017) was a multi-class classification approach.

5.2 Multi-class classification

Table 4 reports our experimental results of a single run with the four different settings[5]. Switching from binary classification to multi-class classification we observe a significant drop in precision and a lower F1 score. This is expected since the classifier now has more TLINK as options from where to decide. Despite of this change, the model keeps outperforming both UTHealth and the state-of-the-art.

5.3 Multi-class classification with PubMed word embeddings

Once we confirmed the adapted model gives competitive results on the narrative container identification task, we focused on increasing the system's recall. Therefore, we changed the word representations for in-domain word embeddings in comparison with the previous experiment, which uses word vectors trained on Wikipedia. Word representation depends on the words in context and because the clinical domain is a very specific field with a different vocabulary of that used in the general domain, we expected the model to benefit from a resource like PubMed. However, our results suggest this does not have a significant impact on most TLINKs (OVERLAP did not change at all). Only BEGINS-ON and ENDS-ON recall considerably improved.

5.4 Multi-class classification with PubMed word embeddings and filtered negative examples

While we increased recall by using in-domain word embeddings, we can still witness an imbalance between precision and recall. Moreover, we are still below UTHealth recall score (highest on CONTAINS identification task). To improve further the model's recall, around 10% of NONE:EVENT-EVENT pairs were removed from the dataset based on a rule of the annotation guidelines that prevents non-real events (i.e. events that do not actually appear on the patient's timeline) to be linked with real events. Recall was further improved for most TLINKs while it remained the same for ENDS-ON. Under this setting, our model reached its best F1 scores for all TLINKs, outperforming the state-of-the-art on CONTAINS.

[5]We experimented a couple of additional runs but the results were always the same.

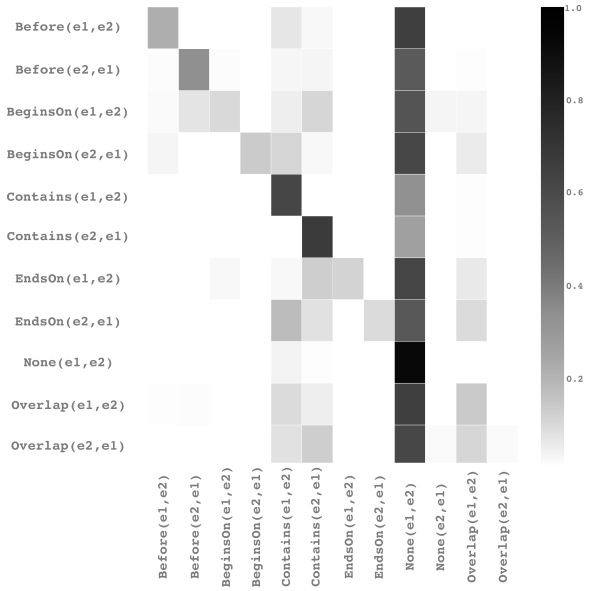

Figure 2: Confusion matrix of our multi-class classification model with PubMed word embedding on the dev set.

6 Error Analysis

We focused our error analysis on the fourth of our experiments. Systems participating in the Clinical TempEval narrative container identification task only received credit if for a pair of entities, 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. Part of our analysis is to determine whether type classification or directionality identification is the most difficult task or if they are both equally problematic for the model. Confusion matrix on Figure 2 shows 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. At the same time, we can observe that 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. 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 0.024. Not even the performance on $Before(e_2, e_1)$ with 0.34 is as low, even though

True relation	Predicted relation	Sentence
$Overlap(e_1, e_2)$	$Contains(e_1, e_2)$	1. **Tumor** *invades* into the muscularis propria.
$Overlap(e_1, e_2)$	$Contains(e_1, e_2)$	2. Recurrent rectal **adenocarcinoma**, previously *resected* node-negative
$Overlap(e_1, e_2)$	$Contains(e_1, e_2)$	3. June 30, 2009: Due to change in stool, patient underwent **colonoscopy** *noting* mass in the right colon.
$Overlap(e_1, e_2)$	$Contains(e_1, e_2)$	4. A **biopsy** obtained was positive for *adenocarcinoma*, consistent with colorectal primary and confirmed by LCC.
$Overlap(e_1, e_2)$	$Contains(e_1, e_2)$	5. **Pathology** from the extended right hemicolectomy was positive for invasive moderately differentiated *adenocarcinoma* in the ascending colon.
$Overlap(e_1, e_2)$	$Contains(e_1, e_2)$	6. Exploratory **surgery** with *appendicitis* 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 *hospitalized* 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** *once every several days*.
$Overlap(e_2, e_1)$	$Contains(e_2, e_1)$	9. She is still **moving** her bowels *multiple times a day*.
$Overlap(e_2, e_1)$	$Contains(e_2, e_1)$	10. The patient **smokes** cigars *about once-a-month*.

Table 5: Sample of the analyzed misclassified sentences by our system. e_1 and e_2 are shown in bold and italics, respectively.

they have similar number of instances (290 and 353, respectively). $Overlap(e_1, e_2)$ with 0.14 is comparable to $BeginsOn(e_2, e_1)$, despite of having 7 times more instances (1291 vs. 176). For this reason, 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 126 misclassified sentences while in $Overlap(e_2, e_1)$ there were 37. EVENT-EVENT pairs were the predominant type of pair in the former while TIMEX3-EVENT were for the latter, with 116 and 29 instances, respectively. We took all of the aforementioned misclassified sentences for supplementary examination and discuss the reason(s) of this errors in the following section.

6.1 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 since 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, OVERLAP is the only symmetrical TLINK type since e_1 OVERLAP e_2 means the same as e_2 OVERLAP e_1.

Strictly speaking, every entity occupies time. An entity's time interval is crucial for understanding its temporal relation with respect to another entity, specially in the case of CONTAINS and OVERLAP relations where the end point of the target is key to determine whether there is complete containment or not. The temporal relations used by the THYME project rely on Allen (1990) interval algebra, a precise way to express 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 thanks to two important grammatical systems: tense and aspect. It is particularly to our interest the definition of aspect, the means with which speakers discuss a single situation, for example, as beginning, continuation, or completion (Li and Shirai, 2000). One of the best known and widely accepted aspect classifications is that of Vendler, who distinguished four categories for verb and verb phrases: *activities*, *accomplishments*, *achievements* and *states*.

Figure 3 presents Vendler's classification using (Andersen, 1990) schematization. Arrows are

used to 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.

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*

Figure 3: Vendler's four-way classification. Abbreviations: C, Clear; NC, Not Clear

Categorizing the source and target entities of a relation as one of Vendler's types simplifies the TLINK classification task. For example, categories with no clear end points like *activities* and *states* are more likely to overlap with *accomplishments* and *achievements*, which have clear end points. Figure 4 illustrates an OVERLAP and CONTAINS relations using Allen's and Vendler's representation of time periods. Leveraging on aspectual type for temporal relation extraction is a promising approach that has already been explored by Costa and Branco (2012) on TempEval data. However, this approach is limited since aspect is a property of verbs.

When analyzing OVERLAP relations that were mistaken for CONTAINS, we realized that just a few events are verbs. Events in sentences 1, 3 and 9 in Table 5 are some examples of this ("invades", "noting" and "moving"). This pointed out the necessity of discriminating between verbal and nonverbal events to understand how they are temporally related. Our observations suggest that rather than recognizing an entity semantic type (e.g. sign or symptoms, diseases, procedures) it is imperative to take into account the action associated to it. Thus, procedures like colonoscopy, biopsy, pathology and surgery have to be *performed*, a dynamic verb with a natural end point: an *accomplishment*. Diseases like adenocarcinoma and appendicitis are *present*, they exist, and consequently they fall in the *state* category. Following this line of reasoning, it is easier to differentiate an OVERLAP relation from CONTAINS in sentence 5 since we understand the adenocarcinoma was found during the performance of the pathology but there is not enough information to tell whether the adenocarcinoma 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*, done indeterminably throughout the day as *multiple times a day* imply. In sentence 7, on the other hand, there is a time expression with a definite time interval overlapping the patient's state of being hospitalized.

Temporally locating two events on a timeline requires a high level of reasoning that even for humans can turn into a complicated task. All of the aforementioned inferences were done heavily relying on the internal constituency of an event, implying Costa and Branco (2012) claim that temporal information processing can profit from information about aspectual type is valid in the clinical domain. Due to the high similarity of CONTAINS and OVERLAP relations it does not come as a surprise that these two types are easily confused by our system, which performed reasonably well on identifying other TLINK types with similar number of instances. This suggests than the main problem is not the amount of data available but how temporal properties are encoded in language.

Aspectual information proved useful for differentiating between two of the most frequent and most similar TLINK types: CONTAINS and OVERLAP. As previously mentioned in Section 4.1, there is a contextual aspect attribute available for EVENT entities with three possible values: *N/A* (default), *NOVEL* and *INTERMITTENT*. The latter could be useful to identify an *activity* or an *accomplishment* but 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 it is more common to find nouns as events. We discuss this finding in more detail in the following section.

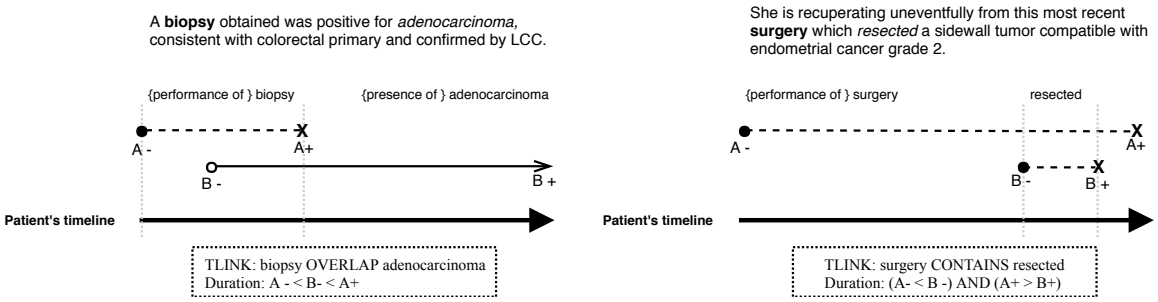

Figure 4: Allen's and Vendler's interval representation of OVERLAP and CONTAINS relations. A- / B- and A+ / B+ represent the start and end of an event, respectively. Filled-dots represent clear start points while an empty-dot represent a not-clear start point.

7 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. A significant amount of OVERLAP links were EVENT-EVENT relations and they also made around half of CONTAINS links. We looked further into these type of pairs, discriminating between verb and non-verbal events. Table 6 shows the results in more detail.

Dev set: Event-Event pairs				
TLINK	**V-V**	**V-NV**	**NV-V**	**NV-NV**
CONTAINS	6	47	24	103
OVERLAP	6	55	27	193
Total	12	102	51	296

Table 6: Distribution of misclassified CONTAINS and OVERLAP Event-Event pairs by type of EVENT. Abbreviations: V, Verb; NV, Non-Verb

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 indicates that most of EVENT-EVENT missclasified pairings were either of NV-NV type or include a NV entity. Time intervals of NV entities like "pain" or "resection" are more difficult to understand, while V entities like "removed" or "improving" have their time properties morphologically encoded. Thus, 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.

The high frequency of NV entities is likely to be one of the reasons why not only our system but also previous works in temporal relation extraction are behind human performance. In the previous section we introduced Vendler's aspectual classification and discussed how it helps separate two extremely similar TLINKs. Unfortunately, this is not compatible with nominal predicates. Verb/Non-Verb entities distinction of EVENTs is a first step that could alleviate this problem and positively influence the temporal relation extraction task.

8 Conclusion and Future work

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 it 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.

Up to now, 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 of

the reasons behind. 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 like the Unified Medical Language System (UMLS) Metathesaurus help to classify entities into semantic types like *Therapeutic or Preventive procedure*, *Sign or Symptom* or *Disease or Syndrome*. Still, the associated events and actions cannot be found in this or any other knowledge base. We hypothesize that a resource containing aspectual information of the actions associated to common nominals like procedures or diseases can further improve temporal relation extraction in the clinical domain. With that in mind, we plan to analyze further EVENT-EVENT relations differentiating events as verbal and non-verbal events.

Acknowledgments

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

References

James F Allen. 1990. Maintaining knowledge about temporal intervals. In *Readings in qualitative reasoning about physical systems*, pages 361–372. Elsevier.

Roger W. Andersen. 1990. Unpublished lecture in the seminar on the acquisition of tense and aspect.

Steven Bethard, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. Semeval-2015 task 6: Clinical tempeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 806–814.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. Semeval-2016 task 12: Clinical tempeval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1052–1062.

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.

Veera Raghavendra Chikka. 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)*, pages 1237–1240.

Francisco Costa and António Branco. 2012. Aspectual type and temporal relation classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 266–275. Association for Computational Linguistics.

Dmitriy Dligach, Timothy Miller, Chen Lin, Steven Bethard, and Guergana Savova. 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*, volume 2, pages 746–751.

Jason Alan Fries. 2016. Brundlefly at semeval-2016 task 12: Recurrent neural networks vs. joint inference for clinical temporal information extraction.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 94–99.

Hee-Jin Lee, Hua Xu, Jingqi Wang, Yaoyun Zhang, Sungrim Moon, Jun Xu, and Yonghui Wu. 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)*, pages 1292–1297.

Artuur Leeuwenberg and Marie-Francine Moens. 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)*, pages 1280–1285.

Artuur Leeuwenberg and Marie-Francine Moens. 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*.

Peng Li and Heng Huang. 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)*, pages 1268–1273.

Ping Li and Yasuhiro Shirai. 2000. *The acquisition of lexical and grammatical aspect*, volume 16. Walter de Gruyter.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2016. Improving temporal relation extraction with training instance augmentation. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 108–113.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR abs/1301.3781*.

Makoto Miwa and Mohit Bansal. 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)*, pages 1105–1116. Association for Computational Linguistics.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. Iso-timeml: An international standard for semantic annotation. In *LREC*, volume 10, pages 394–397.

James Pustejovsky and Amber Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 152–160.

Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. 2011. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39(5):952.

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

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 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)*, volume 2, pages 1–9.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th international workshop on semantic evaluations*, pages 75–80.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 57–62.