SemEval-2017 Task 12: Clinical TempEval

Steven Bethard
University of Arizona
Tucson, AZ 85721, USA
bethard@email.arizona.edu

Guergana Savova
Harvard Medical School
Boston, MA 02115, USA
Guergana.Savova@childrens.harvard.edu

Martha Palmer
University of Colorado Boulder
Boulder, CO 80309
martha.palmer@colorado.edu

James Pustejovsky
Brandeis University
Waltham, MA 02453, USA
jamesp@cs.brandeis.edu

Abstract

Clinical TempEval 2017 aimed to answer the question: how well do systems trained on annotated timelines for one medical condition (colon cancer) perform in predicting timelines on another medical condition (brain cancer)? Nine sub-tasks were included, covering problems in time expression identification, event expression identification and temporal relation identification. Participant systems were evaluated on clinical and pathology notes from Mayo Clinic cancer patients, annotated with an extension of TimeML for the clinical domain. 11 teams participated in the tasks, with the best systems achieving F1 scores above 0.55 for time expressions, above 0.70 for event expressions, and above 0.30 for temporal relations. Most tasks observed about a 20 point drop over Clinical TempEval 2016, where systems were trained and evaluated on the same domain (colon cancer).

1 Introduction

The TempEval shared tasks have, since 2007, provided a focus for research on temporal information extraction (Verhagen et al., 2007, 2010; UzZaman et al., 2013). In recent years the community has moved toward testing such information extraction systems on clinical data, to address a common need of doctors and clinical researchers to search over timelines of clinical events like symptoms, diseases, and procedures. In the Clinical TempEval shared tasks (Bethard et al., 2015, 2016), participant systems have competed to identify critical components of the timeline of a clinical text: time expressions, event expressions, and temporal relations. For example, Figure 1 shows the annotations that a system is expected to produce when given the text:

> April 23, 2014: The patient did not have any postoperative bleeding so we'll resume chemotherapy with a larger bolus on Friday even if there is slight nausea.

Clinical TempEval 2017 introduced a new aspect to this problem: domain adaptation. Whereas in Clinical TempEval 2015 and 2016, systems were both trained and tested on notes from colon cancer patients, in 2017, systems were trained on colon cancer patients, but tested on brain cancer patients. The diseases, symptoms, procedures, etc. vary widely across these two patient populations, and the doctors treating these different kinds of cancer make a variety of different linguistic choices when discussing such patients. As a result, systems that participated in Clinical TempEval 2017 were faced with a much more challenging task than systems from 2015 or 2016.

2 Data

The Clinical TempEval corpus was based on a set of clinical notes and pathology reports from 200 colon cancer patients and 200 brain cancer patients at the Mayo Clinic. These notes were manually de-identified by the Mayo Clinic to replace names, locations, etc. with generic placeholders, but time expressions were not altered. The notes were then manually annotated by the THYME project (thyme.healthnlp.org) using an extension of ISO-TimeML for the annotation of times, events and temporal relations in clinical notes (Styler, IV et al., 2014b). This extension includes additions such as new time expression types (e.g., PREPOSTEXP for expressions like *postoperative*), new EVENT attributes (e.g., DEGREE=LITTLE for expressions like *slight nausea*), and an increased focus on temporal relations of type CONTAINS (a.k.a. INCLUDES).

The annotation procedure was as follows:

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 565–572,
Vancouver, Canada, August 3 - 4, 2017.

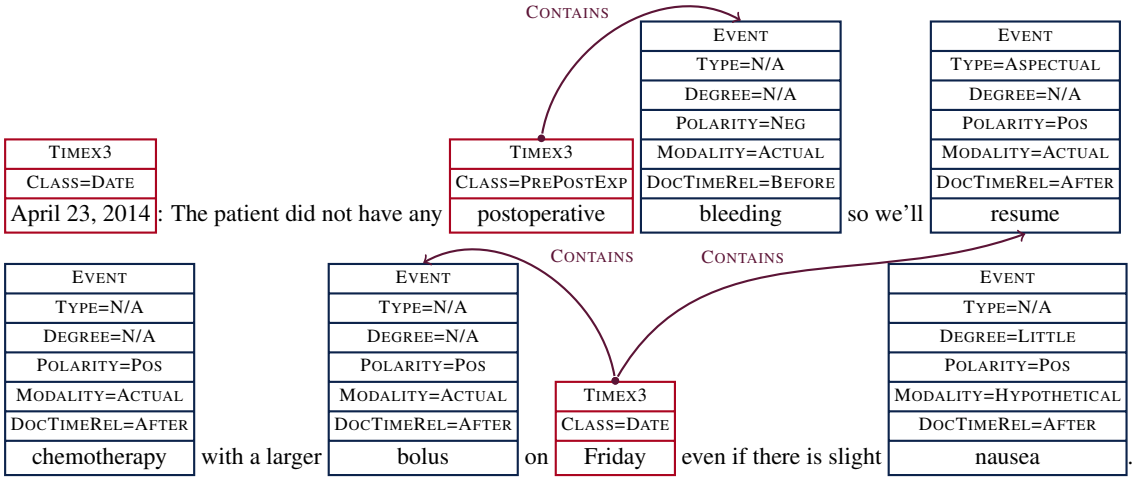

Figure 1: Example Clinical TempEval annotations

1. Annotators identified time and event expressions, along with their attributes
2. Adjudicators revised and finalized the time and event expressions and their attributes
3. Annotators identified temporal relations between pairs of events and events and times
4. Adjudicators revised and finalized the temporal relations

More details on the corpus annotation process are documented in Styler, IV et al. (2014a).

Because the data contained incompletely de-identified clinical data (the time expressions were retained), participants were required to sign a data use agreement with the Mayo Clinic to obtain the raw text of the clinical notes and pathology reports.[1] The event, time and temporal relation annotations were distributed separately from the text, in an open source repository[2] using the Anafora standoff format (Chen and Styler, 2013).

Each corpus (colon cancer and brain cancer) was split into three portions: Train (50%), Dev (25%) and Test (25%). Patients were sorted by patient number (an integer arbitrarily assigned by the de-identification process) and stratified across these splits. Table 1 shows the number of documents, event expressions (EVENT annotations), time expressions (TIMEX3 annotations) and narrative container relations (TLINK annotations with TYPE=CONTAINS attributes) in the Train, Dev, and Test portions of each corpus.

The raw text of both the colon cancer and brain cancer corpora were already released as part of Clinical TempEval 2015 and 2016, as were the time, event, and temporal relation annotations for the colon cancer corpus. However, none of the annotations for the brain cancer corpus were previously released.

Clinical TempEval 2017 ran several phases of evaluation, where different data were released for training and testing sets[3].

Trial This phase replicated the Clinical TempEval 2016 setup: systems were expected to train on the colon cancer Train and Dev sets, and were tested on the colon cancer Test set. This phase was organized primarily to allow participants to validate the format of their system output.

Unsupervised Domain Adaptation In this phase, systems were expected to train on all the colon cancer annotations (released in previous Clinical TempEvals) and were tested on the annotations of the brain cancer Test set. No brain cancer annotations were provided for training, though systems were free to use the raw brain cancer text if they had a way to do so.

Supervised Domain Adaptation This phase released annotations for the first 10 patients in the brain cancer Train data (Train-10 in Table 1). Systems were expected to train on these brain cancer annotations, in addition to the colon cancer annotations provided previously, and were

[1] Details on the data use agreement process can be found at: http://thyme.healthnlp.org/

[2] https://github.com/stylerw/thymedata

[3] All releases were made at the CodaLab site: https://competitions.codalab.org/competitions/15621

	Colon Cancer			Brain Cancer			
	Train	Dev	Test	Train-10	Train	Dev	Test
Documents	293	147	151	30	298	149	148
TIMEX3s	3833	2078	1952	350	3527	1498	1552
EVENTS	38890	20974	18990	2557	26210	11162	11510
TLINKS with TYPE=CONTAINS	11150	6163	5894	624	3938	1641	1759

Table 1: Number of documents, event expressions, time expressions and narrative container relations in Train, Dev, and Test portions of the THYME data. All colon cancer data was released as part of Clinical TempEval 2015 and 2016. The Train-10 column is the data from the first 10 patients of the brain cancer Train data, which was the only additional training data released in Clinical TempEval 2017.

tested on the annotations of the brain cancer Test set. Systems were again free to use all the raw brain cancer text if they had a way to do so.

Note that across all phases, the only brain cancer data released was the Train-10 set. The remainder of the brain cancer data was reserved for future evaluations.

3 Tasks

Nine tasks were included (the same as those of Clinical TempEval 2015 and 2016), grouped into three categories:

- Identifying time expressions (TIMEX3 annotations in the THYME corpus) consisting of the following components:
  - The span (character offsets) of the expression in the text
  - Class: DATE, TIME, DURATION, QUANTIFIER, PREPOSTEXP, or SET
- Identifying event expressions (EVENT annotations in the THYME corpus) consisting of the following components:
  - The span (character offsets) of the expression in the text
  - Contextual Modality: ACTUAL, HYPOTHETICAL, HEDGED, or GENERIC
  - Degree: MOST, LITTLE, or N/A
  - Polarity: POS or NEG
  - Type: ASPECTUAL, EVIDENTIAL, or N/A
- Identifying temporal relations between events and times, focusing on the following types:
  - Relations between events and the document creation time (BEFORE, OVERLAP, BEFORE-OVERLAP, or AFTER), represented by DOCTIMEREL annotations.
  - Narrative container relations (Pustejovsky and Stubbs, 2011), which indicate that an event or time is temporally contained in (i.e., occurred during) another event or time, represented by TLINK annotations with TYPE=CONTAINS.

4 Evaluation Metrics

All of the tasks were evaluated using the standard metrics of precision (P), recall (R) and F_1:

$$P = \frac{|S \cap H|}{|S|} \quad R = \frac{|S \cap H|}{|H|} \quad F_1 = \frac{2 \cdot P \cdot R}{P + R}$$

where S is the set of items predicted by the system and H is the set of items annotated by the humans. Applying these metrics only requires a definition of what is considered an "item" for each task.

- For evaluating the spans of event expressions or time expressions, items were tuples of (begin, end) character offsets. Thus, systems only received credit for identifying events and times with exactly the same character offsets as the manually annotated ones.
- For evaluating the attributes of event expressions or time expressions – Class, Contextual Modality, Degree, Polarity and Type – items were tuples of (begin, end, value) where begin and end are character offsets and value is the value that was given to the relevant attribute. Thus, systems only received credit for an event (or time) attribute if they both found an event (or time) with the correct character offsets and then assigned the correct value for that attribute.
- For relations between events and the document creation time, items were tuples of (begin, end, value), just as if it were an event attribute. Thus, systems only received credit if they found a correct event and assigned the correct relation

(BEFORE, OVERLAP, BEFORE-OVERLAP, or AFTER) between that event and the document creation time.

- For narrative container relations, items were tuples of ((begin$_1$, end$_1$), (begin$_2$, end$_2$)), where the begins and ends corresponded to the character offsets of the events or times participating in the relation. Thus, systems only received credit for a narrative container relation if they found both events/times and correctly assigned a CONTAINS relation between them.

For narrative container relations, the P and R definitions were modified to take into account *temporal closure*, where additional relations are deterministically inferred from other relations (e.g., A CONTAINS B and B CONTAINS C, so A CONTAINS C):

$$P = \frac{|S \cap \text{closure}(H)|}{|S|} \quad R = \frac{|\text{closure}(S) \cap H|}{|H|}$$

Similar measures were used in prior work (UzZaman and Allen, 2011) and TempEval 2013 (UzZaman et al., 2013), following the intuition that precision should measure the fraction of system-predicted relations that can be verified from the human annotations (either the original human annotations or annotations inferred from those through closure), and that recall should measure the fraction of human-annotated relations that can be verified from the system output (either the original system predictions or predictions inferred from those through closure).

5 Human Agreement

We also provide two types of human agreement on the tasks, measured with the same evaluation metrics as the systems:

ann-ann Inter-annotator agreement between the two independent human annotators who annotated each document. This is the most commonly reported type of agreement, and often considered to be an upper bound on system performance.

adj-ann Inter-annotator agreement between the adjudicator and the two independent annotators. This is usually a better bound on system performance in adjudicated corpora, since the models are trained on the adjudicated data, not on the individual annotator data.

Only F_1 is reported in these scenarios since precision and recall depend on the arbitrary choice of one annotator as human (H) and the other as system (S).

6 Baseline Systems

Two rule-based systems were used as baselines to compare the participating systems against.

memorize For all tasks but the narrative container task, a memorization baseline was used.

To train the model, all phrases annotated as either events or times in the training data were collected. All exact character matches for these phrases in the training data were then examined, and only phrases that were annotated as events or times greater than 50% of the time were retained. For each phrase, the most frequently annotated type (event or time) and attribute values for instances of that phrase were determined.

To predict with the model, the raw text of the test data was searched for all exact character matches of any of the memorized phrases, preferring longer phrases when multiple matches overlapped. Wherever a phrase match was found, an event or time with the memorized (most frequent) attribute values was predicted.

closest For the narrative container task, a proximity baseline was used. Each time expression was predicted to be a narrative container, containing only the closest event expression to it in the text.

7 Participating Systems

11 teams submitted a total of 28 runs, 10 for the unsupervised domain adaptation phase, and 18 for the supervised domain adaptation phase. All participating systems trained some form of supervised classifiers, with common features including character n-grams, words, part-of-speech tags, and Unified Medical Language System (UMLS) concept types. Below is a brief description of each participating team, and a note if they performed any more elaborate domain adaptation than simply adding the extra 30 brain cancer notes to their training data.

GUIR (MacAvaney et al., 2017) combined conditional random fields, rules, and decision tree ensembles, with features including character n-grams, words, word shapes, word clusters, word embeddings, part-of-speech tags, syntactic

and dependency tree paths, semantic roles, and UMLS concept types.

Hitachi (P R et al., 2017) combined conditional random fields, rules, neural networks, and decision tree ensembles, with features including character n-grams, word n-grams, word shapes, word embeddings, verb tense, section headers, and sentence embeddings.

KULeuven-LIIR (Leeuwenberg and Moens, 2017) combined support vector machines and structured perceptrons with features including words and part-of-speech tags. For domain adaptation, KULeuven-LIIR tried assigning higher weight to the brain cancer training data, and representing unknown words in the input vocabulary.

LIMSI-COT (Tourille et al., 2017) combined recurrent neural networks with character and word embeddings, and support vector machines with features including words and part-of-speech tags. For domain adaptation, LIMSI-COT tried disallowing modification of pre-trained word embeddings, and representing unknown words in the input vocabulary.

NTU-1 (Huang et al., 2017) combined support vector machines and conditional random fields with features including word n-grams, part-of-speech tags, word shapes, named entities, dependency trees, and UMLS concept types.

ULISBOA (Lamurias et al., 2017) combined conditional random fields and rules with features including character n-grams, words, part-of-speech tags, and UMLS concept types.

XJNLP (Long et al., 2017) combined rules, support vector machines, and recurrent and convolutional neural networks, with features including words, word embeddings, and verb tense.

Several other teams (WuHanNLP, UNICA, UTD, and IIIT) also competed, but did not submit a system description.

8 Evaluation Results

Tables 2 to 4 show the results of the evaluation. In all tables, the best system score from each column is in bold. Systems marked with † were submitted after the competition deadline, and are thus not considered part of the official evaluation.

Team	time span			time span + class		
	F1	P	R	F1	P	R
	Unsupervised domain adaptation					
GUIR	**0.57**	0.61	0.53	0.51	0.55	0.47
KULeuven-LIIR	0.56	**0.72**	0.46	**0.53**	**0.68**	0.43
LIMSI-COT	0.51	0.42	**0.66**	0.49	0.40	**0.63**
ULISBOA	0.48	0.44	0.54	0.43	0.39	0.48
Hitachi	0.43	0.63	0.33	-	-	-
baseline	0.36	**0.72**	0.24	0.32	0.63	0.21
WuHanNLP	0.31	0.65	0.20	0.27	0.57	0.18
	Supervised domain adaptation					
GUIR	**0.59**	0.57	0.62	**0.56**	0.54	0.59
LIMSI-COT	0.58	0.51	**0.67**	0.55	0.49	**0.64**
NTU-1	0.58	**0.58**	0.58	0.54	0.54	0.54
KULeuven-LIIR	0.56	0.57	0.55	0.54	**0.55**	0.53
ULISBOA	0.55	0.52	0.60	0.52	0.48	0.56
UTD	0.54	0.56	0.52	0.44	0.46	0.43
Hitachi	0.51	0.53	0.48	-	-	-
WuHanNLP	0.43	0.45	0.41	0.40	0.42	0.38
XJNLP†	0.41	0.33	0.52	0.35	0.29	0.45
UNICA	0.37	0.31	0.45	0.31	0.26	0.38
baseline	0.35	0.53	0.26	0.32	0.49	0.24
IIIT	0.31	0.39	0.25	0.19	0.24	0.16
	Annotator agreement					
ann-ann	0.81	-	-	0.79	-	-
adj-ann	0.86	-	-	0.85	-	-

Table 2: System performance and annotator agreement on TIMEX3 tasks: identifying the time expression's span (character offsets) and class (DATE, TIME, DURATION, QUANTIFIER, PREPOSTEXP or SET).

8.1 Time Expressions

Table 2 shows results on the time expression tasks. The GUIR system had the top F1 in almost all time expression tasks across both unsupervised and supervised domain adaptation phases, achieving F1s between 0.51 and 0.59. Compared to human agreement, the best systems were more than 0.20 lower than the inter-annotator agreement (and further, of course, from the annotator-adjudicator agreement).

In Clinical TempEval 2016, for comparison, when models were both trained and tested on colon cancer notes, the top system achieved 0.80 F1 for time spans, and 0.77 F1 for time types. This suggests that a time expression system trained on one clinical condition (e.g., colon cancer) can expect a 20+ point drop when tested on another clinical condition (e.g., brain cancer). Providing 30 annotated notes in the target domain narrowed that gap by only a few points.

The drop in performance can probably be partly attributed to differences in time expressions across the two corpora. For example, *post-op* is 26.5 times more common in brain cancer (212 occurrences in brain cancer data vs. 27 occurrences in colon cancer data), *overnight* is 13 times more common (148

Team	event span			event span + modality			event span + degree			event span + polarity			event span + type		
	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R
	Unsupervised domain adaptation														
LIMSI-COT	**0.72**	0.62	**0.84**	**0.64**	0.55	**0.75**	**0.71**	0.62	**0.83**	**0.69**	0.60	**0.82**	**0.70**	0.61	**0.82**
GUIR	0.71	0.64	0.80	0.56	0.50	0.64	0.68	0.61	0.77	0.65	0.59	0.74	0.68	0.61	0.76
KULeuven-LIIR	0.68	**0.70**	0.67	0.62	**0.63**	0.61	0.67	**0.69**	0.66	0.67	**0.68**	0.65	0.66	**0.67**	0.65
ULISBOA	0.68	0.62	0.77	0.61	0.55	0.68	0.68	0.61	0.76	0.66	0.60	0.74	0.66	0.60	0.74
Hitachi	0.68	0.67	0.69	-	-	-	-	-	-	-	-	-	-	-	-
baseline	0.63	0.65	0.61	0.55	0.57	0.54	0.62	0.64	0.60	0.58	0.60	0.56	0.60	0.62	0.59
WuHanNLP	0.62	0.59	0.66	0.55	0.52	0.58	0.61	0.58	0.65	0.6	0.57	0.63	0.60	0.57	0.63
	Supervised domain adaptation														
LIMSI-COT	**0.76**	**0.69**	0.85	**0.69**	**0.63**	**0.78**	**0.75**	**0.68**	0.84	**0.75**	**0.68**	0.83	**0.75**	**0.68**	0.83
GUIR	0.74	0.68	0.82	0.66	0.60	0.72	0.73	0.67	0.80	0.58	0.54	0.64	0.72	0.66	0.79
NTU-1	0.73	0.62	**0.87**	0.63	0.54	0.75	0.72	0.62	**0.86**	0.70	0.60	**0.84**	0.70	0.60	**0.85**
ULISBOA	0.73	0.65	0.83	0.64	0.57	0.73	0.72	0.64	0.82	0.71	0.63	0.81	0.71	0.63	0.80
KULeuven-LIIR	0.72	0.67	0.78	0.66	0.61	0.71	0.71	0.66	0.77	0.71	0.66	0.76	0.70	0.65	0.76
Hitachi	0.71	0.67	0.76	-	-	-	-	-	-	-	-	-	-	-	-
baseline	0.70	0.67	0.74	0.62	0.59	0.65	0.69	0.66	0.73	0.66	0.62	0.69	0.68	0.65	0.72
UTD	0.66	0.62	0.71	0.57	0.53	0.61	-	-	-	-	-	-	-	-	-
WuHanNLP	0.65	0.59	0.72	0.58	0.53	0.64	0.64	0.58	0.71	0.63	0.57	0.70	0.63	0.57	0.70
IIIT	0.62	**0.69**	0.56	0.51	0.57	0.47	0.61	0.67	0.55	0.58	0.64	0.52	0.59	0.66	0.54
XJNLP†	0.61	0.55	0.68	0.51	0.46	0.57	0.59	0.54	0.67	0.54	0.49	0.61	0.58	0.52	0.66
UNICA	0.50	0.39	0.71	0.43	0.34	0.61	0.49	0.38	0.70	0.47	0.37	0.66	0.47	0.37	0.67
	Annotator agreement														
ann-ann	0.79	-	-	0.72	-	-	0.78	-	-	0.78	-	-	0.76	-	-
adj-ann	0.87	-	-	0.84	-	-	0.86	-	-	0.86	-	-	0.85	-	-

Table 3: System performance and annotator agreement on EVENT tasks: identifying the event expression's span (character offsets), contextual modality (ACTUAL, HYPOTHETICAL, HEDGED or GENERIC), degree (MOST, LITTLE or N/A), polarity (POS or NEG) and type (ASPECTUAL, EVIDENTIAL or N/A).

in brain vs. 11 in colon), and *intraoperative* is 2.3 times more common (156 in brain vs. 68 in colon). Formatting is also different across the corpora. For example, *POST-OP* (all capitals) occurs 161 times in all the brain cancer data, but never occurs with this capitalization in any of the colon cancer data.

8.2 Event Expressions

Table 3 shows results on the event expression tasks. The LIMSI-COT system achieved the best F1 on all event expression tasks for both the unsupervised and supervised domain adaptation phases, achieving around 0.70 F1 for most subtasks in the unsupervised setting, and around 0.75 F1 in the supervised setting. Compared to human agreement, the LIMSI-COT system ranged between 0.06 and 0.09 below the inter-annotator agreement.

In Clinical TempEval 2016, for comparison, the top system achieved F1s of 0.92, 0.87, 0.91, 0.90, and 0.89 for event spans, modality, degree, polarity, and type, respectively. This suggests that, much like for time expressions, an event expression system trained on one clinical condition (e.g., colon cancer) can expect a 20+ point drop when tested on another clinical condition (e.g., brain cancer). Providing 30 annotated notes in the target domain again narrows the gap by only a few points.

The drop in performance can again probably be attributed to differences across the two corpora. Even more so than time expressions, event expressions for brain cancer are very different from event expressions for colon cancer. For example, *craniotomy*, *glioma*, *glioblastoma*, *oligoastrocytoma*, *aphasia*, and *temozolomide* all occur as events more than 150 times in the brain cancer data, but do not occur as events even once in the colon cancer data.

8.3 Temporal Relations

Table 4 shows performance on the temporal relation tasks. The LIMSI-COT system had the top F1 in almost all of the temporal relation tasks in both the unsupervised and supervised domain adaptation settings, achieving above 0.50 F1 in linking events to the document creation time, and above 0.30 F1 for linking events to their narrative containers. Compared to humans, the LIMSI-COT system was more than 0.30 below inter-annotator agreement for narrative container relations, but above inter-annotator agreement (though still below annotator-adjudicator agreement) on document time relations when using the additional target domain (brain cancer) training data.

In Clinical TempEval 2016, for comparison, the top system achieved F1s of 0.76 for document time

	To document time			Narrative containers		
	F1	P	R	F1	P	R
	Unsupervised domain adaptation					
LIMSI-COT	**0.51**	0.44	**0.60**	0.33	0.28	**0.40**
KULeuven-LIIR	0.49	**0.50**	0.48	0.32	0.33	0.30
GUIR	0.40	0.36	0.45	**0.34**	**0.52**	0.25
Hitachi	0.45	0.44	0.45	0.23	0.23	0.22
baseline	0.38	0.39	0.37	0.14	0.39	0.08
ULISBOA	0.41	0.37	0.45	-	-	-
WuHanNLP	0.41	0.39	0.43	-	-	-
	Supervised domain adaptation					
LIMSI-COT	**0.59**	**0.53**	**0.66**	**0.32**	0.25	**0.43**
KULeuven-LIIR	0.56	0.52	0.61	0.28	0.23	0.35
GUIR	0.50	0.45	0.55	0.25	**0.59**	0.16
NTU-1	0.49	0.42	0.59	0.26	0.20	0.37
Hitachi	0.52	0.49	0.55	0.16	0.11	0.27
baseline	0.46	0.43	0.48	0.14	0.27	0.09
WuHanNLP	0.46	0.42	0.51	0.12	0.16	0.09
UTD	0.45	0.42	0.48	0.11	0.08	0.16
ULISBOA	0.44	0.39	0.51	-	-	-
IIIT	0.36	0.40	0.33	0.05	0.03	0.08
UNICA	0.20	0.15	0.28	-	-	-
	Annotator agreement					
ann-ann	0.52	-	-	0.66	-	-
adj-ann	0.71	-	-	0.80	-	-

Table 4: System performance and annotator agreement on temporal relation tasks: identifying relations between events and the document creation time (DOCTIMEREL), and identifying narrative container relations (CONTAINS).

relations, and 0.48 for narrative containers. Again we see a major drop when training on one condition (e.g., colon cancer) and testing on another (e.g., brain cancer): a 20+ point drop for document time relations, and around a 15 point drop for narrative containers.

9 Discussion

Clinical TempEval 2017 showed that developing clinical timeline extraction tools that generalize across domains is still a challenging problem. Almost across the board, we saw 20+ point drops in performance when systems were trained on one domain (colon cancer) and tested on another (brain cancer), as compared to systems that were trained and tested on a single domain (colon cancer, as in Clinical TempEval 2016). And across the board, providing a small amount of target domain (brain cancer) training data narrowed that gap only by a couple of points. This is an important finding because it stresses how much work remains to build robust clinical information extraction tools that are useful across a wide range of medical applications.

Though the focus in Clinical TempEval 2017 was on domain adaptation, only a small number of fairly simple domain adaptation techniques were applied by participants, probably because producing even an initial system for all the Clinical TempEval sub-tasks is already a significant effort. Two participants (LIMSI-COT and KULeuven-LIIR, two of the top ranking systems) included special handling of unknown words to try to increase generalization power. Other approaches attempted by participants included giving a heavier weight to the target domain (brain cancer) training data, and using pre-trained domain independent word embeddings. A wide variety of more sophisticated domain adaptation techniques exist that were not applied by participants, and we expect that some of these will make future progress in reducing the cross-domain performance degradation that was observed in Clinical TempEval 2017.

Acknowledgements

The project described was supported in part by R01LM010090 (THYME) from the National Library Of Medicine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

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)*. Association for Computational Linguistics, Denver, Colorado, pages 806–814. http://www.aclweb.org/anthology/S15-2136.

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)*. Association for Computational Linguistics, San Diego, California, pages 1052–1062. http://www.aclweb.org/anthology/S16-1165.

Wei-Te Chen and Will Styler. 2013. Anafora: A web-based general purpose annotation tool. In *Proceedings of the 2013 NAACL HLT Demonstration Session*. Association for Computational Linguistics, Atlanta, Georgia, pages 14–19. http://www.aclweb.org/anthology/N13-3004.

Po-Yu Huang, Hen-Hsen Huang, Yu-Wun Wang, Ching Huang, and Hsin-Hsi Chen. 2017. NTU-1 at SemEval-2017 Task 12: Detection and classification of temporal events in clinical data with domain adaptation. In *Proceedings of the 11th International*

Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics, Vancouver, Canada, pages 1007–1010. http://www.aclweb.org/anthology/S17-2177.

Andre Lamurias, Diana Sousa, Sofia Pereira, Luka Clarke, and Francisco M Couto. 2017. ULISBOA at SemEval-2017 Task 12: Extraction and classification of temporal expressions and events. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 1016–1020. http://www.aclweb.org/anthology/S17-2179.

Artuur Leeuwenberg and Marie-Francine Moens. 2017. KULeuven-LIIR at SemEval-2017 Task 12: Cross-domain temporal information extraction from clinical records. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 1027–1031. http://www.aclweb.org/anthology/S17-2181.

Yu Long, Zhijing Li, Xuan Wang, and Chen Li. 2017. XJNLP at SemEval-2017 Task 12: Clinical temporal information extraction with a hybrid model. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 1011–1015. http://www.aclweb.org/anthology/S17-2178.

Sean MacAvaney, Arman Cohan, and Nazli Goharian. 2017. GUIR at SemEval-2017 Task 12: A framework for cross-domain clinical temporal information extraction. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 1021–1026. http://www.aclweb.org/anthology/S17-2180.

Sarath P R, Manikandan R, and Yoshiki Niwa. 2017. Hitachi at SemEval-2017 Task 12: System for temporal information extraction from clinical notes. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 1002–1006. http://www.aclweb.org/anthology/S17-2176.

James Pustejovsky and Amber Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*. Association for Computational Linguistics, Portland, Oregon, USA, pages 152–160. http://www.aclweb.org/anthology/W11-0419.

William F. Styler, IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014a. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics* 2:143–154. https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/305.

William F. Styler, IV, Guergana Savova, Martha Palmer, James Pustejovsky, Tim O'Gorman, and Piet C. de Groen. 2014b. THYME annotation guidelines. http://clear.colorado.edu/compsem/documents/THYMEGuidelines.pdf.

Julien Tourille, Olivier Ferret, Xavier Tannier, and Aurélie Névéol. 2017. LIMSI-COT at SemEval-2017 Task 12: Neural architecture for temporal information extraction from clinical narratives. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 595–600. http://www.aclweb.org/anthology/S17-2098.

Naushad UzZaman and James Allen. 2011. Temporal evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, pages 351–356. http://www.aclweb.org/anthology/P11-2061.

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)*. Association for Computational Linguistics, Atlanta, Georgia, USA, pages 1–9. http://www.aclweb.org/anthology/S13-2001.

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 Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics, Prague, Czech Republic, pages 75–80. http://www.aclweb.org/anthology/S/S07/S07-1014.

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*. Association for Computational Linguistics, Uppsala, Sweden, pages 57–62. http://www.aclweb.org/anthology/S10-1010.