SemEval-2016 Task 12: Clinical TempEval

Steven Bethard
University of Alabama at Birmingham
Birmingham, AL 35294, USA
bethard@cis.uab.edu

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

Wei-Te Chen
University of Colorado Boulder
Boulder, CO 80309
weite.chen@colorado.edu

Leon Derczynski
University of Sheffield
Sheffield, S1 4DP, UK
leon.derczynski@sheffield.ac.uk

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

Marc Verhagen
Brandeis University
Waltham, MA 02453, USA
marc@cs.brandeis.edu

Abstract

Clinical TempEval 2016 evaluated temporal information extraction systems on the clinical domain. Nine sub-tasks were included, covering problems in time expression identification, event expression identification and temporal relation identification. Participant systems were trained and evaluated on a corpus of clinical and pathology notes from the Mayo Clinic, annotated with an extension of TimeML for the clinical domain. 14 teams submitted a total of 40 system runs, with the best systems achieving near-human performance on identifying events and times. On identifying temporal relations, there was a gap between the best systems and human performance, but the gap was less than half the gap of Clinical TempEval 2015.

1 Introduction

The TempEval shared tasks have, since 2007, provided a focus for research on temporal information extraction (Verhagen et al., 2007; Verhagen et al., 2010; UzZaman et al., 2013). Participant systems compete to identify critical components of the timeline of a text, including time expressions, event expressions and temporal relations. However, the TempEval campaigns to date have focused primarily on in-document timelines derived from news articles. In recent years, the community has moved toward testing such information extraction systems on clinical data (Sun et al., 2013; Bethard et al., 2015) to broaden our understanding of the language of time beyond newswire expressions and structure.

Clinical TempEval focuses on discrete, well-defined tasks which allow rapid, reliable and repeatable evaluation. Participating systems are expected to take as input raw text, for example:

> 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.

The systems are then expected to output annotations over the text, for example, those shown in Figure 1. That is, the systems should identify the time expressions, event expressions, attributes of those expressions, and temporal relations between them.

Clinical TempEval 2016 addressed one of the major challenges in Clinical TempEval 2015: data distribution. Because Clinical TempEval is based on real patient notes from the Mayo Clinic, participants go through a lengthy authorization process involving a data use agreement and an interview. For Clinical TempEval 2016, we streamlined this process and were able to authorize data access for more than twice as many participants as Clinical TempEval 2015. And since all the training and evaluation data distributed for Clinical TempEval 2015 was used as the training data for Clinical TempEval 2016, participants had more than a year to work on their systems. The result was that four times as many teams participated.

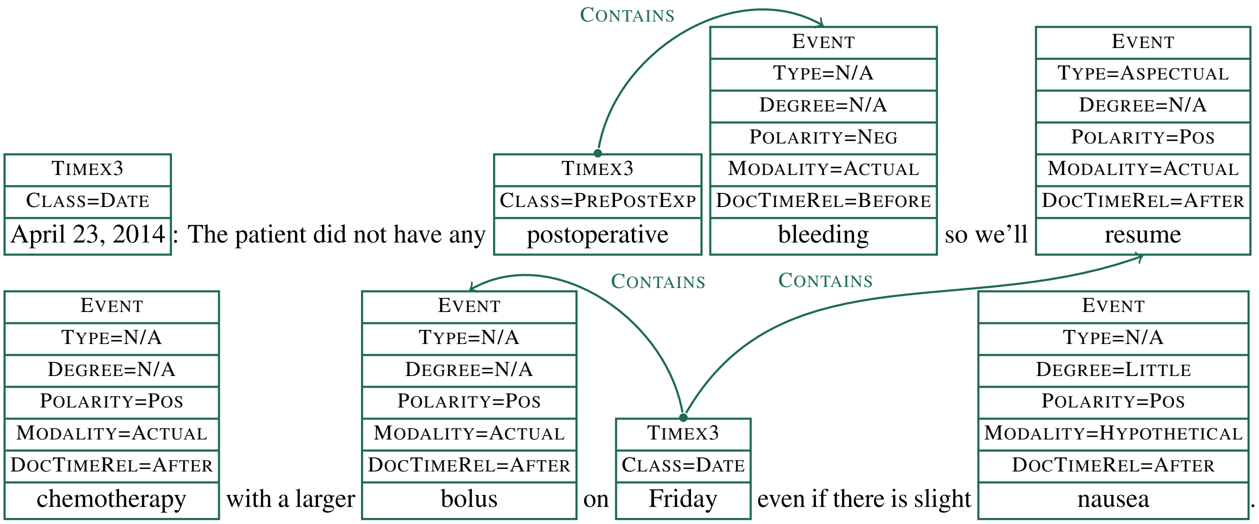

Figure 1: Example Clinical TempEval annotations

2 Data

The Clinical TempEval corpus was based on a set of 600 clinical notes and pathology reports from 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 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:

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 a separate article (Styler 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).

The corpus 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. The Train and Dev portions were released to participants for training and tuning their systems. The Test portion was reserved for evaluation of the systems. 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 the corpus.

3 Tasks

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

- Identifying time expressions (TIMEX3 annotations in the THYME corpus) consisting of the

[1] Details on the process: `http://thyme.healthnlp.org/`

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

	Train	Dev	Test
Documents	293	147	151
Timex3s	3833	2078	1952
Events	38890	20974	18990
Type=Contains Tlinks	11176	6173	5894

Table 1: Number of documents, event expressions, time expressions and narrative container relations in Train, Dev, and Test portions of the THYME data. Both Train and Dev were released as part of Clinical TempEval 2015.

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.

The evaluation was run in two phases:

1. Systems were provided access only to the raw text, and were asked to identify time expressions, event expressions and temporal relations
2. Systems were provided access to the raw text and the manual event and time annotations, and were asked to identify only temporal relations

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. In the second phase of the evaluation, when manual event annotations were provided as input, only recall (which in this case is equivalent to standard classification accuracy) is reported.
- 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 event and time attributes, we also measure how accurately a system predicts the attribute values on just those events or times that the system predicted. The goal here is to allow a comparison across systems for assigning attribute values, even when different systems produce different numbers of events and times. This metric is calculated by dividing the F_1 on the attribute by the F_1 on identifying the spans:

$$A = \frac{\text{attribute } F_1}{\text{span } F_1}$$

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 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.

6 Participating Systems

14 research teams submitted a total of 40 runs:

brundlefly (Fries, 2016) submitted 1 run for phase 1 based on recurrent neural networks, word embeddings, and logistic regression, and 1 run for phase 2 run based on the DeepDive framework (`http://deepdive.stanford.edu`).

CDE-IIITH (Chikka, 2016) submitted 2 runs for each phase, the first based on deep learning models, and the second based on conditional random fields and support vector machines.

Cental (Hansart et al., 2016) submitted 1 run for phase 1, based on conditional random fields and lexical resources.

GUIR (Cohan et al., 2016) submitted 2 runs for phase 1 and 1 run for phase 2, based on conditional random fields and logistic regression with lexical, morphological, syntactic, dependency, and domain specific features, combined with pattern matching rules.

HITACHI (Sarath P R et al., 2016) submitted 2 runs for the time portion of phase 1, based on ensembles of rule-based and machine learning systems with lexical, syntactic and morphological features. The second run included 50% more training data than the first.

KULeuven-LIIR (Leeuwenberg and Moens, 2016) submitted 2 runs for phase 2, based on the cTAKES-temporal machine-learning model (Lin et al., 2015), with additional features.

LIMSI (Grouin and Moriceau, 2016) submitted 2 runs for each phase, based on conditional random fields with lexical, morphological, and word cluster features, and the rule-based HeidelTime (Strötgen and Gertz, 2013).

LIMSI-COT (Tourille et al., 2016) submitted 2 runs for phase 2, the first based on support vector machines with lexical, syntactic, structural, and UMLS features, and the second based on replacing the lexical features with word embeddings.

ULISBOA (Barros et al., 2016) submitted 2 runs for each phase, based on the IBEnt framework's support vector machines with lexical and morphological features (`https://github.com/AndreLamurias/IBEnt`), and rule-based extensions to Stanford CoreNLP (Manning et al., 2014). The runs differed on how rules were incorporated for each subtask.

UtahBMI (AAl Abdulsalam et al., 2016) submitted 2 runs for each phase, the first based on conditional random fields and the second based on support vector machines. Both runs used lexical, morphological, syntactic, shape, character pattern, character n-gram, section type, and gazetteer features.

UTA-MLNLP (Li and Huang, 2016) submitted 2 runs for each phase, based on a neural network with a different window size for each run.

UTHealth (Lee et al., 2016) submitted 2 runs for each phase, based on linear and structural (HMM) support vector machines using lexical, morphological, syntactic, discourse, and word representation features. The runs differed on the features included.

VUACLTL (Caselli and Morante, 2016) submitted 2 runs for each phase, based on conditional random fields with morpho-syntactic, lexical, UMLS, and DBpedia features. The first run was a two-step approach to temporal relations, the second, a one step approach.

7 Human Agreement

We also provide two types of human agreement on the task, 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.

Precision and recall are not reported in these scenarios since they depend on the arbitrary choice of one annotator as human (H) and the other as system (S).

Note that since temporal relations between events and the document creation time were annotated at the same time as the events themselves, agreement for this task is only reported in phase 1 of the evaluation. Similarly, since narrative container relations were only annotated after events and times had been adjudicated, agreement for this task is only reported in phase 2 of the evaluation.

8 Evaluation Results

8.1 Time Expressions

Table 2 shows results on the time expression tasks. The UTHealth systems achieved the best results on almost all time-related tasks. For finding times, while one system had comparable precision to UTHealth (0.836 UTHealth vs. 0.840 LIMSI), no system had competitive recall (0.757 UTHealth vs. 0.714 from the next best, UtahBMI), and thus the UTHealth system consistently outperformed the other systems in F_1. The results were similar for jointly finding times and assigning them a time class, though a couple systems (HITACHI, GUIR) did have more accurate predictions for the time class when scored only on the times that they were able to find (0.971 UTHealth vs. 0.975 HITACHI vs. 0.989 GUIR).

Compared to human agreement, the UTHealth and UtahBMI systems exceeded the inter-annotator agreement on times of 0.731, but even UTHealth's F_1 of 0.795 did not reach the annotator-adjudicator agreement of 0.830, and the results were similar for jointly finding times and assigning their classes (0.772 vs. 0.807). Nonetheless, these 0.025 and 0.035 gaps between the top system and the human agreement are smaller than the 0.051 and 0.038 gaps observed in Clinical TempEval 2015 (Bethard et al., 2015).

8.2 Event Expressions

Table 3 shows results on the event expression tasks. Again, UTHealth dominated the field, achieving the highest score on almost every event-related task. However, the gap to the second place team was much smaller for events than it was for times: only a 0.011

	span			span + class			
Team	P	R	F1	P	R	F1	A
UTHealth-1	0.836	**0.757**	**0.795**	0.812	**0.735**	**0.772**	0.971
UTHealth-2	0.826	0.758	0.790	0.800	0.734	0.765	0.968
UtahBMI-crf	0.798	0.714	0.754	0.771	0.690	0.729	0.967
UtahBMI-svm	0.810	0.690	0.745	0.792	0.674	0.728	0.977
HITACHI-1	0.781	0.685	0.730	0.759	0.671	0.712	0.975
HITACHI-2	0.781	0.668	0.720	0.758	0.654	0.702	0.975
Cental-crf	0.777	0.564	0.653	0.752	0.545	0.632	0.968
LIMSI-2	0.830	0.518	0.638	0.804	0.503	0.619	0.970
LIMSI-1	**0.840**	0.510	0.635	**0.815**	0.495	0.616	0.970
CDE-IIITH-crf	0.752	0.515	0.612	0.644	0.439	0.522	0.853
CDE-IIITH-dl	0.614	0.560	0.586	0.468	0.426	0.446	0.761
brundlefly	0.686	0.415	0.517	0.639	0.387	0.482	0.932
VUACLTL-1	0.660	0.372	0.476	0.638	0.363	0.462	0.971
VUACLTL-2	0.660	0.372	0.476	0.638	0.363	0.462	0.971
GUIR-2	0.649	0.256	0.367	0.640	0.253	0.362	0.986
GUIR-1	0.486	0.273	0.349	0.480	0.269	0.345	**0.989**
Baseline: memorize	0.774	0.428	0.551	0.746	0.413	0.532	0.966
GUIR†	0.802	0.678	0.735	0.775	0.655	0.710	0.966
ULISBOA-2†	0.776	0.692	0.732	-	-	-	-
ULISBOA-1†	0.623	0.065	0.118	-	-	-	-
Agreement: ann-ann	-	-	0.731	-	-	0.688	0.941
Agreement: adj-ann	-	-	0.830	-	-	0.807	0.972

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). The best system score from each column is in bold. Systems marked with † were submitted after the competition deadline and are not considered official.

gap between UTHealth's 0.903 F_1 and UtahBMI's 0.892. The gap was even smaller if we look at precision and recall separately: a 0.007 gap between UTHealth's 0.915 precision and UTA's 0.908, and a 0.005 gap between UTHealth's 0.891 precision and UtahBMI's 0.886. The results were similar for most of the attributes, though the precision gaps were larger (1.1-1.4) and the recall gaps were smaller (0.3-0.7).

Compared to human agreement, UTHealth, UtahBMI, Cental, GUIR, and UTA all exceeded inter-annotator agreement on identifying events, and UTHealth and UtahBMI exceeded inter-annotator agreement on all of the attributes. None of the systems reached the level of annotator-adjudicator agreement: even UTHealth's F_1 on events of 0.903 had a gap of 0.019 from the annotator-adjudicator agreement of 0.922, and the results were similar for event attributes: 0.049 for modality, 0.021 for degree, 0.029 for polarity, 0.024 for type. These gaps are almost all bigger than the gaps observed in Clinical TempEval 2015: 0.005 for event spans, 0.031 for modality, 0.007 for degree, 0.012 for polarity, 0.030 for type. However, Clinical TempEval 2016's human agreement was substantially higher, with all annotator-adjudicator agreement above 0.90, while in Clinical TempEval 2015, annotator-adjudicator agreement ranged from 0.853 to 0.880.

8.3 Temporal Relations

Table 4 shows performance on the temporal relation tasks. In both phase 1 (where systems were provided only the raw text) and phase 2 (where systems were provided the manually annotated events and times), the UTHealth system was again the top system for most tasks. For relating events to the document creation time, the UTHealth system had the best precision, recall, and F_1 (0.766, 0.746, and 0.756) in phase

	span			span + modality				span + degree			
Team	P	R	F1	P	R	F1	A	P	R	F1	A
UTHealth-1	**0.915**	**0.891**	**0.903**	**0.866**	**0.843**	**0.855**	**0.947**	**0.911**	**0.887**	**0.899**	0.996
UTHealth-2	0.903	0.886	0.895	0.855	0.839	0.847	0.946	0.899	0.883	0.891	0.996
UtahBMI-svm	0.897	0.886	0.892	0.841	0.831	0.836	0.937	0.892	0.881	0.887	0.994
UtahBMI-crf	0.902	0.883	0.892	0.850	0.832	0.841	0.943	0.898	0.879	0.889	**0.997**
Cental-crf	0.892	0.878	0.885	-	-	-	-	-	-	-	-
GUIR-2	0.887	0.872	0.880	0.836	0.822	0.829	0.942	0.883	0.868	0.875	0.994
GUIR-1	0.886	0.872	0.879	0.830	0.817	0.824	0.937	0.882	0.868	0.875	0.995
UTA-4	0.908	0.842	0.874	0.842	0.780	0.810	0.927	0.904	0.838	0.869	0.994
UTA-5	0.900	0.850	0.874	0.837	0.790	0.813	0.930	0.896	0.845	0.870	0.995
VUACLTL-1	0.868	0.828	0.847	0.795	0.758	0.776	0.916	0.864	0.824	0.844	0.996
VUACLTL-2	0.868	0.828	0.847	0.795	0.758	0.776	0.916	0.864	0.824	0.844	0.996
LIMSI-1	0.885	0.808	0.845	0.811	0.742	0.775	0.917	0.880	0.805	0.841	0.995
LIMSI-2	0.869	0.816	0.842	0.798	0.749	0.772	0.917	0.865	0.812	0.838	0.995
CDE-IIITH-crf	0.835	0.797	0.815	0.764	0.729	0.746	0.915	0.830	0.793	0.811	0.995
CDE-IIITH-dl	0.838	0.786	0.811	0.779	0.731	0.754	0.930	0.834	0.783	0.807	0.995
brundlefly	0.883	0.660	0.755	0.819	0.612	0.701	0.928	0.878	0.657	0.752	0.996
Baseline: memorize	0.878	0.834	0.855	0.810	0.770	0.789	0.923	0.874	0.831	0.852	0.996
GUIR†	0.891	0.872	0.881	0.836	0.818	0.827	0.939	0.887	0.868	0.877	0.995
ULISBOA-1†	0.881	0.745	0.807	-	-	-	-	-	-	-	-
ULISBOA-2†	0.879	0.739	0.803	-	-	-	-	-	-	-	-
Agreement: ann-ann	-	-	0.864	-	-	0.833	0.964	-	-	0.861	0.997
Agreement: adj-ann	-	-	0.922	-	-	0.904	0.980	-	-	0.920	0.998

	span + polarity				span + type			
Team	P	R	F1	A	P	R	F1	A
UTHealth-1	**0.900**	**0.875**	**0.887**	0.982	**0.894**	**0.870**	**0.882**	**0.977**
UTHealth-2	0.888	0.872	0.880	**0.983**	0.880	0.863	0.871	0.973
UtahBMI-svm	0.879	0.869	0.874	0.980	0.854	0.843	0.849	0.952
UtahBMI-crf	0.885	0.867	0.876	0.982	0.875	0.857	0.866	0.971
Cental-crf	0.870	0.857	0.864	0.976	-	-	-	-
GUIR-2	0.871	0.856	0.864	0.982	0.864	0.850	0.857	0.974
GUIR-1	0.869	0.855	0.862	0.981	0.863	0.850	0.857	0.975
UTA-4	0.876	0.812	0.842	0.963	0.877	0.813	0.844	0.966
UTA-5	0.861	0.813	0.836	0.957	0.869	0.820	0.844	0.966
VUACLTL-1	0.780	0.743	0.761	0.898	0.839	0.800	0.819	0.967
VUACLTL-2	0.780	0.743	0.761	0.898	0.839	0.800	0.819	0.967
LIMSI-1	0.867	0.792	0.828	0.980	0.825	0.754	0.788	0.933
LIMSI-2	0.851	0.799	0.824	0.979	0.811	0.761	0.785	0.932
CDE-IIITH-crf	0.750	0.716	0.733	0.899	0.806	0.769	0.787	0.966
CDE-IIITH-dl	0.813	0.764	0.788	0.972	0.814	0.765	0.789	0.973
brundlefly	0.856	0.640	0.733	0.971	0.829	0.620	0.709	0.939
Baseline: memorize	0.812	0.772	0.792	0.926	0.855	0.813	0.833	0.974
GUIR†	0.875	0.856	0.866	0.983	0.868	0.849	0.858	0.974
Agreement: ann-ann	-	-	0.852	0.986	-	-	0.835	0.966
Agreement: adj-ann	-	-	0.916	0.993	-	-	0.906	0.983

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). The best system score from each column is in bold. Systems marked with † were submitted after the competition deadline and are not considered official.

	To document time			Narrative containers		
	P	R	F1	P	R	F1
Phase 1: systems are provided only the raw text						
UTHealth-1	**0.766**	**0.746**	**0.756**	0.488	**0.471**	**0.479**
UTHealth-2	0.757	0.743	0.750	0.479	0.466	0.472
UtahBMI-crf	0.753	0.737	0.745	0.502	0.215	0.301
UtahBMI-svm	0.741	0.732	0.736	0.498	0.215	0.300
GUIR-2	0.719	0.707	0.713	-	-	-
GUIR-1	0.712	0.701	0.706	-	-	-
VUACLTL-1	0.655	0.624	0.639	**0.531**	0.244	0.334
VUACLTL-2	0.655	0.624	0.639	0.493	0.268	0.347
CDE-IIITH-dl	0.643	0.604	0.623	0.285	0.225	0.252
LIMSI-1	0.635	0.580	0.607	-	-	-
LIMSI-2	0.624	0.585	0.604	-	-	-
CDE-IIITH-crf	0.481	0.460	0.470	0.431	0.167	0.241
brundlefly	0.389	0.290	0.332	-	-	-
UTA-4	0.340	0.315	0.327	-	-	-
UTA-5	0.336	0.317	0.326	-	-	-
Baseline: memorize / closest	0.620	0.589	0.604	0.403	0.067	0.115
GUIR†	0.719	0.704	0.711	-	-	-
ULISBOA-1†	-	-	-	0.122	0.009	0.017
ULISBOA-2†	-	-	-	0.108	0.009	0.017
Agreement: ann-ann	-	-	0.721	-	-	-
Agreement: adj-ann	-	-	0.844	-	-	-
Phase 2: systems are provided manually annotated EVENTs and TIMEX3s						
UTHealth-1	-	0.835	-	0.588	**0.559**	**0.573**
UTHealth-2	-	0.833	-	0.568	0.564	0.566
LIMSI-COT-lexical	-	0.769	-	0.704	0.436	0.538
GUIR-1	-	0.813	-	0.546	0.471	0.506
LIMSI-COT-embedding	-	0.807	-	**0.751**	0.320	0.449
KULeuven-LIIR-1	-	-	-	0.714	0.428	0.536
KULeuven-LIIR-2	-	-	-	0.715	0.429	0.536
VUACLTL-2	-	0.701	-	0.589	0.368	0.453
VUACLTL-1	-	0.701	-	0.642	0.345	0.449
UtahBMI-crf+svm	-	**0.843**	-	0.562	0.254	0.350
CDE-IIITH-dl	-	0.705	-	0.348	0.284	0.313
UtahBMI-svm	-	0.571	-	0.605	0.230	0.333
ULISBOA-1	-	-	-	0.273	0.255	0.264
brundlefly	-	0.742	-	-	-	-
uta-5	-	0.788	-	-	-	-
uta-6	-	0.786	-	-	-	-
LIMSI-1	-	0.687	-	-	-	-
CDE-IIITH-crf	-	0.588	-	0.493	0.185	0.269
LIMSI-2	-	0.679	-	-	-	-
ULISBOA-2	-	-	-	0.823	0.056	0.105
Baseline: memorize / closest	-	0.675	-	0.459	0.154	0.231
UtahBMI-crf+svm†	-	0.843	-	0.693	0.425	0.527
UtahBMI-svm†	-	0.571	-	0.711	0.372	0.489
Agreement: ann-ann	-	-	-	-	-	0.651
Agreement: adj-ann	-	-	-	-	-	0.817

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). The best system score from each column is in bold. Systems marked with † were submitted after the competition deadline and are not considered official.

1, and the second best score (0.835 vs. UtahBMI's 0.843) in phase 2. For finding narrative container relations, the UTHealth system had the best recall (0.471 in phase 1, 0.559 in phase 2), and though other systems (UtahBMI, VUACLTL, LIMSI-COT, and KULeuven-LIIR) had higher precisions, the recall gap from UTHealth to the next system was large (0.203 in phase 1 and 0.088 in phase 2) and thus UTHealth had the best F_1 in both phases (0.479 in phase 1, 0.573 in phase 2).

Compared to human agreement, UTHealth and UtahBMI exceeded inter-annotator agreement on relations to the document time (while still leaving a gap of 0.088 to the annotator-adjudicator agreement), but no participant system was near the human agreement for narrative containers (a gap of 0.078 from inter-annotator agreement and a gap of 0.244 from annotator-adjudicator agreement). For relations to the document time, the 0.088 gap between systems and annotator-adjudicator agreement is slightly larger than the 0.059 of Clinical TempEval 2015, but for narrative container relations the 0.244 gap is much smaller than the 0.412 of Clinical TempEval 2015. As with other tasks, human agreement is higher this year (0.844 and 0.817 in 2016 vs. 0.761 and 0.672 in 2015), which may explain the larger gap for document time relations. The smaller gap for narrative container relations despite the increased human agreement suggests that major improvements have been made to the systems for this task.

9 Discussion

The results of Clinical TempEval 2016 suggest that current state-of-the-art systems are close to solving most event and time related tasks. For all of these tasks, the gap between system performance and human performance was less than 0.05, and for half the tasks (time spans, event spans, event degree, event type) it was 0.025 or less.

The temporal relation tasks were more difficult. Systems trying to predict the temporal relation between an event and the time at which the document was written lagged about 0.09 behind human performance. And systems trying to predict narrative containers (whether one event or time contains another) lagged about 0.25 behind human performance, even when provided human-annotated events and times. Nonetheless, the latter result was a major improvement over Clinical TempEval 2015, where the gap on narrative containers was more than 0.4.

While there was variability across the subtasks in the rankings of teams, UTHealth and UtahBMI were always at the top of the lists. Both of these systems relied on structured learning models (UTHealth used HMM support vector machines; UtahBMI used conditional random fields) with a wide variety of features (lexical, morphological, syntactic, and many others). We can thus infer that such approaches hold promise for temporal information extraction. However, these two teams were also among the first to make it through the data use agreement process, so their success may in part reflect the advantage of having more time for experimentation and feature engineering on the training data.

Overall, Clinical TempEval 2016 represented a major step forward from Clinical TempEval 2015. It saw a much greater breadth of participating systems (14 teams in 2016 vs. 3 teams in 2015), with the top systems maintaining 2015's high performance on the event and time tasks, while making major progress on the harder temporal relation tasks. Future plans for Clinical TempEval target the robustness of these systems: instead of testing on only colon cancer notes from the Mayo Clinic (the same domain as the training set), systems will be tested on other types of medical conditions and notes from other institutions.

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

Abdulrahman Khalifa AAl Abdulsalam, Sumithra Velupillai, and Stephane Meystre. 2016. UtahBMI at SemEval-2016 Task 12: Extracting temporal information from clinical text. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1256–1262, San Diego, California, June. Association for Computational Linguistics.

Marcia Barros, André Lamúrias, Gonçalo Figueiró, Marta Antunes, Joana Teixeira, Alexandre Pinheiro, and Francisco Couto. 2016. ULISBOA at SemEval-2016 Task

12: Extractions of temporal expressions, clinical events and relations using IBEnt. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

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, Denver, Colorado, June. Association for Computational Linguistics.

Tommaso Caselli and Roser Morante. 2016. VUACLTL at SemEval-2016 Task 12: A crf pipeline to clinical temporal. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

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

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)*, San Diego, California, June. Association for Computational Linguistics.

Arman Cohan, Kevin Meurer, and Nazli Goharian. 2016. GUIR at SemEval-2016 Task 12: Temporal information extraction from clinical narratives. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

Jason Fries. 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)*, San Diego, California, June. Association for Computational Linguistics.

Cyril Grouin and Véronique Moriceau. 2016. LIMSI at SemEval-2016 Task 12: machine-learning and temporal information to identify clinical events and time expressions. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

Charlotte Hansart, Damien De Meyere, Patrick Watrin, André Bittar, and Cédrick Fairon. 2016. Cental at SemEval-2016 Task 12: A linguistically fed CRF model for medical and temporal information extraction. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

Hee-Jin Lee, Hua Xu, Jingqi Wang, Yaoyun Zhang, Sungrim Moon, Jun Xu, and Yonghui Wu. 2016. UTHealth at SemEval-2016 Task 12: Temporal information extraction from clinical notes - uthealth's system for the 2016 clinical tempeval challenge. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

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)*, San Diego, California, June. Association for Computational Linguistics.

Peng Li and Heng Huang. 2016. UTA_MLNLP at SemEval-2016 Task 12. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

Chen Lin, Dmitriy Dligach, Timothy A Miller, Steven Bethard, and Guergana K Savova. 2015. Multilayered temporal modeling for the clinical domain. *Journal of the American Medical Informatics Association*.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June. Association for Computational Linguistics.

James Pustejovsky and Amber Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 152–160, Portland, Oregon, USA, June. Association for Computational Linguistics.

Sarath P R, Manikandan R, and Yoshiki Niwa. 2016. Hitachi at SemEval-2016 Task 12: Extraction of temporal information for the 2016 clinical tempeval challenge. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

Jannik Strötgen and Michael Gertz. 2013. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298.

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.

William F. Styler, IV, Guergana Savova, Martha Palmer, James Pustejovsky, Tim O'Gorman, and Piet C. de Groen. 2014b. THYME annotation guidelines, 2.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

Julien Tourille, Olivier Ferret, Aurélie Névéol, and Xavier Tannier. 2016. LIMSI-COT at SemEval-2016 Task 12: Temporal relation identification for the clinical tempeval challenge. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

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

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)*, pages 1–9, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

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)*, pages 75–80, Prague, Czech Republic, June. Association for Computational Linguistics.

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, Uppsala, Sweden, July. Association for Computational Linguistics.