SemEval-2013 Task 1: TEMPEVAL-3: Evaluating Time Expressions, Events, and Temporal Relations

Naushad UzZaman[♣⋈], **Hector Llorens**[♢⋈], **Leon Derczynski**[♡], **Marc Verhagen**[♠], **James Allen**[♣] and **James Pustejovsky**[♠]

[♣]: University of Rochester, USA; [♢]: University of Alicante, Spain

[♡]: Department of Computer Science, University of Sheffield, UK

[♠]: Computer Science Department, Brandeis University, USA

[⋈]: Nuance Communications

naushad@cs.rochester.edu, hllorens@dlsi.ua.es, leon@dcs.shef.ac.uk

Abstract

Within the SemEval-2013 evaluation exercise, the TempEval-3 shared task aims to advance research on temporal information processing. It follows on from TempEval-1 and -2, with: a three-part structure covering temporal expression, event, and temporal relation extraction; a larger dataset; and new single measures to rank systems – in each task and in general. In this paper, we describe the participants' approaches, results, and the observations from the results, which may guide future research in this area.

1 Introduction

The TempEval task (Verhagen et al., 2009) was added as a new task in SemEval-2007. The ultimate aim of research in this area is the automatic identification of temporal expressions (timexes), events, and temporal relations within a text as specified in TimeML annotation (Pustejovsky et al., 2005). However, since addressing this aim in a first evaluation challenge was deemed too difficult a staged approach was suggested.

TempEval (henceforth TempEval-1) was an initial evaluation exercise focusing only on the categorization of temporal relations and only in English. It included three relation types: event-timex, event-dct,[1] and relations between main events in consecutive sentences.

TempEval-2 (Verhagen et al., 2010) extended TempEval-1, growing into a multilingual task, and consisting of six subtasks rather than three. This included event and timex extraction, as well as the three relation tasks from TempEval-1, with the addition of a relation task where one event subordinates another.

TempEval-3 (UzZaman et al., 2012b) is a follow-up to TempEval 1 and 2, covering English and Spanish. TempEval-3 is different from its predecessors in a few respects:

Size of the corpus: the dataset used has about 600K word silver standard data and about 100K word gold standard data for training, compared to around 50K word corpus used in TempEval 1 and 2. Temporal annotation is a time-consuming task for humans, which has limited the size of annotated data in previous TempEval exercises. Current systems, however, are performing close to the inter-annotator reliability, which suggests that larger corpora could be built from automatically annotated data with minor human reviews. We want to explore whether there is value in adding a large automatically created silver standard to a hand-crafted gold standard.

End-to-end temporal relation processing task: the temporal relation classification tasks are performed from raw text, i.e. participants need to extract their own events and temporal expressions first, determine which ones to link and then obtain the relation types. In previous TempEvals, gold timexes, events, and relations (without category) were given to participants.

Temporal relation types: the full set of temporal relations in TimeML are used, rather than the reduced set used in earlier TempEvals.

Platinum test set: A new test dataset has been developed for this edition. It is based on manual annotations by experts over new text (unseen in previous editions).

Evaluation: we report a temporal awareness score for evaluating temporal relations, which helps to rank systems with a single score.

2 Data

In TempEval-3, we reviewed and corrected existing corpora, and also released new corpora.

2.1 Reviewing Existing Corpora

We considered the existing TimeBank (Pustejovsky et al., 2003) and AQUAINT[2] data for TempEval-3. TempEval-

[1] DCT stands for document creation time

[2] See http://timeml.org/site/timebank/timebank.html

Entity	Agreement
Event	0.87
Event class	0.92
Timex	0.87
Timex value	0.88

Table 1: Platinum corpus entity inter-annotator agreement.

Corpus	# of words	Standard
TimeBank	61,418	Gold
AQUAINT	33,973	Gold
TempEval-3 Silver	666,309	Silver
TempEval-3 Eval	6,375	Platinum
TimeBank-ES Train	57,977	Gold
TimeBank-ES Eval	9,833	Gold

Table 2: Corpora used in TempEval-3.

1 and TempEval-2 had the same documents as TimeBank but different relation types and events.

For both TimeBank and AQUAINT, we, (i) cleaned up the formatting for all files making it easy to review and read, (ii) made all files XML and TimeML schema compatible, (iii) added some missing events and temporal expressions. In TimeBank, we, (i) borrowed the events from the TempEval-2 corpus and (ii) borrowed the temporal relations from TimeBank corpus, which contains a full set of temporal relations. In AQUAINT, we added the temporal relations between event and DCT (document creation time), which was missing for many documents in that corpus. These existing corpora comprised the high-quality component of our training set.

2.2 New Corpora

We created two new datasets: a small, manually-annotated set over new text (platinum); and a machine-annotated, automatically-merged dataset based on outputs of multiple systems (silver).

The TempEval-3 *platinum* evaluation corpus was annotated/reviewed by the organizers, who are experts in the area. This process used the TimeML Annotation Guidelines v1.2.1 (Saurí et al., 2006). Every file was annotated independently by at least two expert annotators, and a third was dedicated to adjudicating between annotations and merging the final result. Some annotators based their work on TIPSem annotation suggestions (Llorens et al., 2012b). The GATE Annotation Diff tool was used for merging (Cunningham et al., 2013), a custom TimeML validator ensured integrity,[3] and CAVaT (Derczynski and Gaizauskas, 2010) was used to determine various modes of TimeML mis-annotation and inconsistency that are inexpressable via XML schema. Post-exercise, that corpus (TempEval-3 Platinum with around 6K tokens, on completely new text) is released for the community to review and improve.[4] Inter-annotator agreement (measured with F1, as per Hripcsak and Rothschild (2005)) and the number of annotation passes per document were higher than in existing TimeML corpora, hence the name. Details are given in Table 1. Attribute value scores are given based on the agreed entity set. These are for exact matches.

The TempEval-3 *silver* evaluation corpus is a 600K word corpus collected from Gigaword (Parker et al., 2011). We automatically annotated this corpus by TIPSem, TIPSem-B (Llorens et al., 2013) and TRIOS (UzZaman and Allen, 2010). These systems were retrained on the corrected TimeBank and AQUAINT corpus to generate the original TimeML temporal relation set. We then merged these three state-of-the-art system outputs using our merging algorithm (Llorens et al., 2012a). In our selected merged configuration all entities and relations suggested by the best system (TIPSem) are added in the merged output. Suggestions from other systems (TRIOS and TIPSem-B) are added in the merged output, only if they are also supported by another system. The weights considered in our configuration are: TIPSem 0.36, TIPSemB 0.32, TRIOS 0.32.

For Spanish, Spanish TimeBank 1.0 corpus (Saurí and Badia, 2012) wads used. It is the same corpus that was used in TempEval-2, with a major review of entity annotation and an important improvement regarding temporal relation annotation. For TempEval-3, we converted ES-TimeBank link types to the TimeML standard types based on Allen's temporal relations (Allen, 1983).

Table 2 summarizes our released corpora, measured with PTB-scheme tokens as words. All data produced was annotated using a well-defined subset of TimeML, designed for easy processing, and for reduced ambiguity compared to standard TimeML. Participants were encouraged to validate their submissions using a purpose-built tool to ensure that submitted runs were legible. We called this standard TimeML-strict, and release it separately (Derczynski et al., 2013).

3 Tasks

The three main tasks proposed for TempEval-3 focus on TimeML entities and relations:

3.1 Task A (Timex extraction and normalization)

Determine the extent of the timexes in a text as defined by the TimeML TIMEX3 tag. In addition, determine the value of the features TYPE and VALUE. The possible values of TYPE are time, date, duration, and set; VALUE is a normalized value as defined by the TIMEX3 standard.

[3] See https://github.com/hllorens/TimeML-validator

[4] In the ACL data and code repository, reference `ADCR2013T001`. See also https://bitbucket.org/leondz/te3-platinum

3.2 Task B (Event extraction and classification)

Determine the extent of the events in a text as defined by the TimeML EVENT tag and the appropriate CLASS.

3.3 Task ABC (Annotating temporal relations)

This is the ultimate task for evaluating an end-to-end system that goes from raw text to TimeML annotation of entities and links. It entails performing tasks A and B. From raw text extract the temporal entities (events and timexes), identify the pairs of temporal entities that have a temporal link (TLINK) and classify the temporal relation between them. Possible pair of entities that can have a temporal link are: (i) main events of consecutive sentences, (ii) pairs of events in the same sentence, (iii) event and timex in the same sentence and (iv) event and document creation time. In TempEval-3, TimeML relation are used, i.e.: BEFORE, AFTER, INCLUDES, IS-INCLUDED, DURING, SIMULTANEOUS, IMMEDIATELY AFTER, IMMEDIATELY BEFORE, IDENTITY, BEGINS, ENDS, BEGUN-BY and ENDED-BY.

In addition to this main tasks, we also include two extra temporal relation tasks:

Task C (Annotating relations given gold entities) Given the gold entities, identify the pairs of entities that have a temporal link (TLINK) and classify the temporal relations between them.

Task C relation only (Annotating relations given gold entities and related pairs) Given the temporal entities and the pair of entities that have a temporal link, classify the temporal relation between them.

4 Evaluation Metrics

The metrics used to evaluate the participants are:

4.1 Temporal Entity Extraction

To evaluate temporal entities (*events* and *temporal expressions*), we need to evaluate, (i) How many entities are correctly identified, (ii) If the extents for the entities are correctly identified, and (iii) How many entity attributes are correctly identified. We use classical precision and recall for recognition.

How many entities are correctly identified: We evaluate our entities using the entity-based evaluation with the equations below.

$$Precision = \frac{|Sys_{entity} \cap Ref_{entity}|}{|Sys_{entity}|}$$

$$Recall = \frac{|Sys_{entity} \cap Ref_{entity}|}{|Ref_{entity}|}$$

where, Sys_{entity} contains the entities extracted by the system that we want to evaluate, and Ref_{entity} contains the entities from the reference annotation that are being compared.

If the extents for the entities are correctly identified: We compare our entities with both strict match and relaxed match. When there is a exact match between the system entity and gold entity then we call it strict match, e.g. "sunday morning" vs "sunday morning". When there is a overlap between the system entity and gold entity then we call it relaxed match, e.g. "sunday" vs "sunday morning". When there is a relaxed match, we compare the attribute values.

How many entity attributes are correctly identified: We evaluate our entity attributes using the attribute F1-score, which captures how well the system identified both the entity and attribute (attr) together.

Attribute Recall = $\frac{|\{\forall x \mid x \in (Sys_{entity} \cap Ref_{entity}) \wedge Sys_{attr}(x) == Ref_{attr}(x)\}|}{|Ref_{entity}|}$

Attribute Precision = $\frac{|\{\forall x \mid x \in (Sys_{entity} \cap Ref_{entity}) \wedge Sys_{attr}(x) == Ref_{attr}(x)\}|}{|Sys_{entity}|}$

Attribute F1-score = $\frac{2*p*r}{p+r}$

Attribute (Attr) accuracy, precision and recall can be calculated as well from the above information.

Attr Accuracy = Attr F1 / Entity Extraction F1
*Attr R = Attr Accuracy * Entity R*
*Attr P = Attr Accuracy * Entity P*

4.2 Temporal Relation Processing

To evaluate relations, we use the evaluation metric presented by UzZaman and Allen (2011).[5] This metric captures the temporal awareness of an annotation in terms of precision, recall and F1 score. Temporal awareness is defined as the performance of an annotation as identifying and categorizing temporal relations, which implies the correct recognition and classification of the temporal entities involved in the relations. Unlike TempEval-2 relation score, where only categorization is evaluated for relations, this metric evaluates how well pairs of entities are identified, how well the relations are categorized, and how well the events and temporal expressions are extracted.

$$Precision = \frac{|Sys^{-}_{relation} \cap Ref^{+}_{relation}|}{|Sys^{-}_{relation}|}$$

$$Recall = \frac{|Ref^{-}_{relation} \cap Sys^{+}_{relation}|}{|Ref^{-}_{relation}|}$$

where, G^{+} is the closure of graph G and G^{-} is the reduced of graph G, where redundant relations are removed.[6]

We calculate the $Precision$ by checking the number of reduced system relations ($Sys^{-}_{relation}$) that can be verified from the reference annotation temporal closure graph ($Ref^{+}_{relation}$), out of number of temporal relations in the

[5] We used a minor variation of the formula, where we consider the reduced graph instead of all system or reference relations. Details can be found in Chapter 6 of UzZaman (2012).

[6] A relation is redundant if it can be inferred through other relations.

	F1	P	R	strict F1	**value F1**
HeidelTime-t	90.30	93.08	87.68	81.34	**77.61**
HeidelTime-bf	87.31	90.00	84.78	78.36	72.39
HeidelTime-1.2	86.99	89.31	84.78	78.07	72.12
NavyTime-1,2	**90.32**	89.36	**91.30**	79.57	70.97
ManTIME-4	89.66	95.12	84.78	74.33	68.97
ManTIME-6	87.55	98.20	78.99	73.09	68.27
ManTIME-3	87.06	94.87	80.43	69.80	67.45
SUTime	**90.32**	89.36	**91.30**	79.57	67.38
ManTIME-1	87.20	97.32	78.99	70.40	67.20
ManTIME-5	87.20	97.32	78.99	69.60	67.20
ManTIME-2	88.10	97.37	80.43	72.22	66.67
ATT-2	85.25	98.11	75.36	78.69	65.57
ATT-1	85.60	**99.05**	75.36	79.01	65.02
ClearTK-1,2	90.23	93.75	86.96	**82.71**	64.66
JU-CSE	86.38	93.28	80.43	75.49	63.81
KUL	83.67	92.92	76.09	69.32	62.95
KUL-TE3RunABC	82.87	92.04	75.36	73.31	62.15
ClearTK-3,4	87.94	94.96	81.88	77.04	61.48
ATT-3	80.85	97.94	68.84	72.34	60.43
FSS-TimEx	85.06	90.24	80.43	49.04	58.24
TIPSem (TE2)	84.90	97.20	75.36	81.63	65.31

Table 3: Task A - Temporal Expression Performance.

reduced system relations ($Sys^{-}_{relation}$). Similarly, we calculate the $Recall$ by checking the number of reduced reference annotation relations ($Ref^{-}_{relation}$) that can be verified from the system output's temporal closure graph ($Sys^{+}_{relation}$), out of number of temporal relations in the reduced reference annotation ($Ref^{-}_{relation}$).

This metric evaluates Task ABC together. For Task C and Task C - relation only, all the gold annotation entities were provided and then evaluated using the above metric.

Our evaluation toolkit that evaluated TempEval-3 participants is available online.[7]

5 Evaluation Results

The aim of this evaluation is to provide a meaningful report of the performance obtained by the participants in the tasks defined in Section 3.

Furthermore, the results include TIPSem as reference for comparison. This was used as a pre-annotation system in some cases. TIPSem obtained the best results in event processing task in TempEval-2 and offered very competitive results in timex and relation processing. The best timex processing system in TempEval-2 (HeidelTime) is participating in this edition as well, therefore we included TIPSem as a reference in all tasks.

We only report results in main measures. Results are divided by language and shown per task. Detailed scores can be found on the task website.[8]

[7] See http://www.cs.rochester.edu/u/naushad/temporal

[8] See http://www.cs.york.ac.uk/semeval-2013/task1/

5.1 Results for English

5.1.1 Task A: Timexes

We had nine participants and 21 unique runs for temporal expression extraction task, Task A. Table 3 shows the results. Details about participants' approaches can be found in Table 4.

We rank the participants for Task A on the F1 score of most important timex attribute – *Value*. To get the attribute *Value* correct, a system needs to correctly normalise the temporal expression. This score (*Value* F1) captures the performance of extracting the timex and identifying the attribute *Value* together (*Value F1 = Timex F1 * Value Accuracy*).

Participants approached the temporal expression extraction task with rule-engineered methods, machine learning methods and also hybrid methods. For temporal expression normalization (identifying the timex attribute value), all participants used rule-engineered approaches.

Observations: We collected the following observations from the results and from participants' experiments.

Strategy: Competition was close for timex recognition and the best systems all performed within 1% of each other. On our newswire corpus, statistical systems (ClearTK) performed best at strict matching, and rule-engineered system best at relaxed matching (NavyTime, SUTime, HeidelTime).

Strategy: post-processing, on top of machine learning-base temporal expression extraction, provided a statistically significant improvement in both precision and recall (ManTIME).

Data: using the large silver dataset, alone or together with human annotated data, did not give improvements in performance for Task A. Human-annotated gold standard data alone provided the best performance (ManTIME).

Data: TimeBank alone was better than TimeBank and AQUAINT together for Task A (ClearTK).

Features: syntactic and gazetteers did not provide any statistically significant increment of performance with respect to the morphological features alone (ManTIME).

Regarding the two sub-tasks of timex annotation, recognition and interpretation/normalisation, we noticed a shift in the state of the art. While normalisation is currently (and perhaps inherently) done best by rule-engineered systems, recognition is now done well by a variety of methods. Where formerly, rule-engineered timex recognition always outperformed other classes of approach, now it is clear that rule-engineering and machine learning are equally good at timex recognition.

5.1.2 Task B: Events

For event extraction (Task B) we had seven participants and 10 unique runs. The results for this task can be found in Table 6. We rank the participants for TaskB on the F1 score of most important event attribute – *Class*. *Class*

Strategy	System	Training data	Classifier used
Data-driven	ATT-1, 2, 3	TBAQ + TE3Silver	MaxEnt
	ClearTK-1, 2	TimeBank	SVM, Logit
	ClearTK-3, 4	TBAQ	SVM, Logit
	JU-CSE	TBAQ	CRF
	ManTIME-1	TBAQ + TE3Silver	CRF
	ManTIME-3	TBAQ	CRF
	ManTIME-5	TE3Silver	CRF
	Temp : ESAfeature	TBAQ	MaxEnt
	Temp : WordNetfeature	TBAQ	MaxEnt
	TIPSem (TE2)	TBAQ	CRF
Rule-based	FSS-TimEx (EN)	None	None
	FSS-TimEx (ES)	None	None
	HeidelTime-1.2, bf (EN)	None	None
	HeidelTime-t (EN)	TBAQ	None
	HeidelTime (ES)	Gold	None
	NavyTime-1, 2	None	None
	SUTime	None	None
Hybrid	KUL	TBAQ + TE3Silver	Logit + post-processing
	KUL-TE3RunABC	TBAQ +TE3Silver	Logit + post-processing
	ManTIME-2	TBAQ + TE3Silver	CRF + post-processing
	ManTIME-4	TBAQ	CRF + post-processing
	ManTIME-6	TE3Silver	CRF + post-processing

Table 4: Automated approaches for TE3 Timex Extraction

Strategy	System	Training data	Classifier used	Linguistic Knowledge
Data-driven	ATT-1, 2, 3	TBAQ + TE3Silver	MaxEnt	*ms, ss*
	ClearTK-1, 2	TimeBank	SVM, Logit	*ms*
	ClearTK-3, 4	TBAQ	SVM, Logit	*ms*
	JU-CSE	TBAQ	CRF	
	KUL	TBAQ +TE3Silver	Logit	*ms, ls*
	KUL-TE3RunABC	TBAQ +TE3Silver	Logit	*ms, ls*
	NavyTime-1	TBAQ	MaxEnt	*ms, ls*
	NavyTime-2	TimeBank	MaxEnt	*ms, ls*
	Temp : ESAfeature	TBAQ	MaxEnt	*ms, ls, ss*
	Temp : WordNetfeature	TBAQ	MaxEnt	*ms, ls*
	TIPSem (TE2)	TBAQ	CRF/SVM	*ms, ls, ss*
Rule-based	FSS-TimEx (EN)	None	None	*ls, ms*
	FSS-TimEx (ES)	None	None	*ls, ms*

Table 5: Automated approaches for Event Extraction

	F1	P	R	**class F1**
ATT-1	**81.05**	81.44	80.67	**71.88**
ATT-2	80.91	81.02	**80.81**	71.10
KUL	79.32	80.69	77.99	70.17
ATT-3	78.63	**81.95**	75.57	69.55
KUL-TE3RunABC	77.11	77.58	76.64	68.74
ClearTK-3,4	78.81	81.40	76.38	67.87
NavyTime-1	80.30	80.73	79.87	67.48
ClearTK-1,2	77.34	81.86	73.29	65.44
NavyTime-2	79.37	80.52	78.26	64.81
Temp:ESAfeature	68.97	78.33	61.61	54.55
JU-CSE	78.62	80.85	76.51	52.69
Temp:WordNetfeature	63.90	78.90	53.69	50.00
FSS-TimEx	65.06	63.13	67.11	42.94
TIPSem (TE2)	82.89	83.51	82.28	75.59

Table 6: Task B - Event Extraction Performance.

	F1	P	R
ClearTK-2	**30.98**	34.08	**28.40**
ClearTK-1	29.77	**34.49**	26.19
ClearTK-3	28.62	30.94	26.63
ClearTK-4	28.46	29.73	27.29
NavyTime-1	27.28	31.25	24.20
JU-CSE	24.61	19.17	34.36
NavyTime-2	21.99	26.52	18.78
KUL-TE3RunABC	19.01	17.94	20.22
TIPSem (TE2)	42.39	38.79	46.74

Table 7: Task ABC - Temporal Awareness Evaluation (Task C evaluation from raw text).

F1 captures the performance of extracting the event and identifying the attribute *Class* together (*Class F1 = Event F1 * Class Accuracy*).

All the participants except one used machine learning approaches. Details about the participants' approaches and the linguistic knowledge[9] used to solve this problem, and training data, are in Table 5.

Observations: We collected the following observations from the results and from participants' experiments.

Strategy: All the high performing systems for event extraction (Task B) are machine learning-based.

Data: Systems using silver data, along with the human annotated gold standard data, performed very well (top three participants in the task – ATT, KUL, KUL-TE3RunABC). Additionally, TimeBank and AQUAINT together performed better than just TimeBank alone (NavyTime-1, ClearTK-3,4).

Linguistic Features: Semantic features (*ls* and *ss*) have played an important role, since the best systems (TIPSem, ATT1 and KUL) include them. However, these three are not the only systems using semantic features.

[9]Abbreviations used in the table: TBAQ – *TimeBank + AQUAINT corpus* ms – *morphosyntactic information*, e.g. POS, lexical information, morphological information and syntactic parsing related features; ls –*lexical semantic information*, e.g. WordNet synsets; ss – *sentence-level semantic information*, e.g. Semantic Role labels.

	F1	P	R
ClearTK-2	**36.26**	37.32	35.25
ClearTK-4	35.86	35.17	36.57
ClearTK-1	35.19	**37.64**	33.04
UTTime-5	34.90	35.94	33.92
ClearTK-3	34.13	33.27	35.03
NavyTime-1	31.06	35.48	27.62
UTTime-4	28.81	37.41	23.43
JU-CSE	26.41	21.04	35.47
NavyTime-2	25.84	31.10	22.10
KUL-TE3RunABC	24.83	23.35	26.52
UTTime-1	24.65	15.18	**65.64**
UTTime-3	24.28	15.10	61.99
UTTime-2	24.05	14.80	64.20
TIPSem (TE2)	44.25	39.71	49.94

Table 8: Task C - TLINK Identification and Classification.

	F1	P	R
UTTime-1, 4	**56.45**	**55.58**	**57.35**
UTTime-3, 5	54.70	53.85	55.58
UTTime-2	54.26	53.20	55.36
NavyTime-1	46.83	46.59	47.07
NavyTime-2	43.92	43.65	44.20
JU-CSE	34.77	35.07	34.48

Table 9: Task C - relation only: Relation Classification.

5.1.3 Task C: Relation Evaluation

For complete temporal annotation from raw text (Task ABC - Task C from raw text) and for temporal relation only tasks (Task C, Task C relation only), we had five participants in total.

For relation evaluation, we primarily evaluate on Task ABC (Task C from raw text), which requires joint entity extraction, link identification and relation classification. The results for this task can be found in Table 7.

While TIPSem obtained the best results in task ABC, especially in recall, it was used by some annotators to pre-label data. In the interest of rigour and fairness, we separate out this system.

For task C, for provided participants with entities and participants identified: between which entity pairs a relation exists (link identification); and the class of that relation. Results are given in Table 8. We also evaluate the participants on the relation by providing the entities and the links (performance in Table 9) – TIPSem could not be evaluated in this setting since the system is not prepared to do categorization only unless the relations are divided as in TempEval-2. For these Task C related tasks, we had only one new participant, who didn't participate in Task A and B: UTTime.

Identifying which pair of entities to consider for temporal relations is a new task in this TempEval challenge. The participants approached the problems in data-driven, rule-based and also in hybrid ways (Table 10[10]). On

[10]New abbreviation in the table, e-attr – *entity attributes, e.g. event class, tense, aspect, polarity, modality; timex type, value.*

Strategy	System	Training data	Classifier used	Linguistic Knowledge
Data-driven	ClearTK-1	TimeBank	SVM, Logit	*e-attr, ms*
	ClearTK-2	TimeBank + Bethard et al. (2007)	SVM, Logit	*e-attr, ms*
	ClearTK-3	TBAQ	SVM, Logit	*e-attr, ms*
	ClearTK-4	TBAQ + Muller's inferences	SVM, Logit	*e-attr, ms*
	KULRunABC	TBAQ	SVM, Logit	*ms*
Rule-based	JU-CSE	None	None	
	UTTime-1, 2 ,3	None	None	
	TIPSem (TE2)	None	None	*e-attr, ms, ls, ss*
Hybrid	NavyTime-1	TBAQ	MaxEnt	*ms*
	NavyTime-2	TimeBank	MaxEnt	*ms*
	UTTime-4	TBAQ	Logit	*ms, ls, ss*
	UTTime-5	TBAQ + inverse relations	Logit	*ms, ls, ss*

Table 10: Automated approaches for TE3 TLINK Identification

Strategy	System	Training data	Classifier used	Linguistic Knowledge
Data-driven	ClearTK-1	TimeBank	SVM, Logit	*ms, ls*
	ClearTK-2	TimeBank + Bethard et al. (2007)	SVM, Logit	*ms, ls*
	ClearTK-3	TBAQ	SVM, Logit	*ms, ls*
	ClearTK-4	TBAQ + Muller's inferences	SVM, Logit	*ms, ls*
	JU-CSE	TBAQ	CRF	
	KULRunABC	TBAQ	SVM, Logit	*ms*
	NavyTime-1	TBAQ	MaxEnt	*ms, ls*
	NavyTime-2	TimeBank	MaxEnt	*ms, ls*
	UTTime-1,4, 2	TBAQ	Logit	*ms, ls, ss*
	UTTime-3,5	TBAQ + inverse relations	Logit	*ms, ls, ss*
	TIPSem (TE-2)	TBAQ	CRF/SVM	*ms, ls, ss*

Table 11: Automated approaches for Relation Classification

the other hand, all the participants used data-driven approaches for temporal relations (Table 11).

Observations: We collected the following observations from the results and from participants' experiments.

Strategy: For relation classification, all participants used partially or fully machine learning-based systems.

Data: None of the participants implemented their systems training on the silver data. Most of the systems use the combined TimeBank and AQUAINT (TBAQ) corpus.

Data: Adding additional high-quality relations, either Philippe Muller's closure-based inferences or the verb clause relations from Bethard et al. (2007), typically increased recall and the overall performance (ClearTK runs two and four).

Features: Participants mostly used the morphosyntactic and lexical semantic information. The best performing systems from TempEval-2 (TIPSem and TRIOS) additionally used sentence level semantic information. One participant in TempEval-3 (UTTime) also did deep parsing for the sentence level semantic features.

Features: Using more Linguistic knowledge is important for the task, but it is more important to execute it properly. Many systems performed better using less linguistic knowledge. Hence a system (e.g. ClearTK) with basic morphosyntactic features is hard to beat with more semantic features, if not used properly.

	entity extraction				
	strict	relaxed			
	F1	F1	P	R	value
HeidelTime	**85.3**	**90.1**	96.0	84.9	**87.5**
TIPSemB-F	82.6	87.4	93.7	81.9	82.0
FSS-TimEx	49.5	65.2	86.6	52.3	62.7

Table 12: Task A: Temporal Expression (Spanish).

				class	tense	aspect
	F1	P	R	F1	F1	F1
FSS-TimEx	57.6	89.8	42.4	24.9	-	-
TIPSemB-F	**88.8**	91.7	86.0	**57.6**	41.0	36.3

Table 13: Task B: Event Extraction (Spanish).

Classifier: Across the various tasks, ClearTK tried Mallet CRF, Mallet MaxEnt, OpenNLP MaxEnt, and LIBLINEAR (SVMs and logistic regression). They picked the final classifiers by running a grid search over models and parameters on the training data, and for all tasks, a LIBLINEAR model was at least as good as all the other models. As an added bonus, it was way faster to train than most of the other models.

6 Evaluation Results (Spanish)

There were two participants for Spanish. Both participated in task A and only one of them in task B. In this

	F1	P	R
TIPSemB-F	**41.6**	37.8	46.2

Table 14: Task ABC: Temporal Awareness (Spanish).

	entity extraction				attributes	
	strict	relaxed			val	type
	F1	F1	P	R	F1	F1
HeidelTime	86.4	89.8	94.0	85.9	**87.5**	**89.8**
FSS-TimEx	42.1	68.4	86.7	56.5	48.7	65.8
TIPSem	**86.9**	**93.7**	98.8	89.1	75.4	88.0
TIPSemB-F	84.3	89.9	93.0	87.0	82.0	86.5

Table 15: Task A: TempEval-2 test set (Spanish).

case, TIPSemB-Freeling is provided as a state-of-the-art reference covering all the tasks. TIPSemB-Freeling is the Spanish version of TIPSem with the main difference that it does not include semantic roles. Furthermore, it uses Freeling (Padró and Stanilovsky, 2012) to obtain the linguistic features automatically.

Table 12 shows the results obtained for task A. As it can be observed HeidelTime obtains the best results. It improves the previous state-of-the-art results (TIPSemB-F), especially in normalization (value F1).

Table 13 shows the results from event extraction. In this case, the previous state-of-the-art is not improved.

Table 14 only shows the results obtained in temporal awareness by the state-of-the-art system since there were not participants on this task. We observe that TIPSemB-F approach offers competitive results, which is comparable to results obtained in TE3 English test set.

6.1 Comparison with TempEval-2

TempEval-2 Spanish test set is included as a subset of this TempEval-3 test set. We can therefore compare the performance across editions. Furthermore, we can include the full-featured TIPSem (Llorens et al., 2010), which unlike TIPSemB-F used the AnCora (Taulé et al., 2008) corpus annotations as features including semantic roles.

For timexes, as can be seen in Table 15, the original TIPSem obtains better results for timex extraction, which favours the hypothesis that machine learning systems are very well suited for this task (if the training data is sufficiently representative). However, for normalization (value F1), HeidelTime – a rule-engineered system – obtains better results. This indicates that rule-based approaches have the upper hand in this task. TIPSem uses a partly data-driven normalization approach which, given the small amount of training data available, seemed less suited to the task.

				class	tense	aspect
	F1	P	R	F1	F1	F1
FSS-TimEx	59.0	90.3	43.9	24.6	-	-
TIPSemB-F	**90.2**	92.5	88.0	58.6	39.7	38.1
TIPSem	88.2	90.6	85.8	**58.7**	84.9	78.7

Table 16: Task B: TempEval-2 test set (Spanish).

Table 16 shows event extraction performance in TE2 test set. TIPSemB-F and TIPSem obtained a similar performance. TIPSemB-F performed better in extraction and TIPSem better in attribute classification.

7 Conclusion

In this paper, we described the TempEval-3 task within the SemEval 2013 exercise. This task involves identifying temporal expressions (timexes), events and their temporal relations in text. In particular participating systems were required to automatically annotate raw text using TimeML annotation scheme

This is the first time end-to-end systems are evaluated with a new single score (temporal awareness). In TempEval-3 participants had to obtain temporal relations from their own extracted timexes and events which is a very challenging task and was the ultimate evaluation aim of TempEval. It was proposed at TempEval-1 but has not been carried out until this edition.

The newly-introduced silver data proved not so useful for timex extraction or relation classification, but did help with event extraction. The new single-measure helped to rank systems easily.

Future work could investigate temporal annotation in specific applications. Current annotations metrics evaluate relations for entities in the same consecutive sentence. For document-level understanding we need to understand discourse and pragmatic information. Temporal question answering-based evaluation (UzZaman et al., 2012a) can help us to evaluate participants on document level temporal information understanding without creating any additional training data. Also, summarisation, machine translation, and information retrieval need temporal annotation. Application-oriented challenges could further research in these areas.

From a TimeML point of view, we still haven't tackled subordinate relations (TimeML SLINKs), aspectual relations (TimeML ALINKs), or temporal signal annotation (Derczynski and Gaizauskas, 2011). The critical questions of which links to annotate, and whether the current set of temporal relation types are appropriate for linguistic annotation, are still unanswered.

Acknowledgments

We thank the participants – especially Steven Bethard, Jannik Strötgen, Nate Chambers, Oleksandr Kolomiyets, Michele Filannino, Philippe Muller and others – who helped us to improve TempEval-3 with their valuable feedback. The third author also thanks Aarhus University, Denmark who kindly provided facilities.

References

J. F. Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

S. Bethard, J. H. Martin, and S. Klingenstein. 2007. Timelines from text: Identication of syntactic temporal relations. In *Proceedings of IEEE International Conference on Semantic Computing*.

H. Cunningham, V. Tablan, A. Roberts, and K. Bontcheva. 2013. Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. *PLoS computational biology*, 9(2):e1002854.

L. Derczynski and R. Gaizauskas. 2010. Analysing Temporally Annotated Corpora with CAVaT. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 398–404.

L. Derczynski and R. Gaizauskas. 2011. A Corpus-based Study of Temporal Signals. In *Proceedings of the 6th Corpus Linguistics Conference*.

L. Derczynski, H. Llorens, and N. UzZaman. 2013. TimeML-strict: clarifying temporal annotation. *CoRR*, abs/1304.

G. Hripcsak and A. S. Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298.

H. Llorens, E. Saquete, and B. Navarro. 2010. TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 284–291. Association for Computational Linguistics.

H. Llorens, N. UzZaman, and J. Allen. 2012a. Merging Temporal Annotations. In *Proceedings of the TIME Conference*.

H. Llorens, E. Saquete, and B. Navarro-Colorado. 2012b. Automatic system for identifying and categorizing temporal relations in natural language. *International Journal of Intelligent Systems*, 27(7):680–703.

H. Llorens, E. Saquete, and B. Navarro-Colorado. 2013. Applying Semantic Knowledge to the Automatic Processing of Temporal Expressions and Events in Natural Language. *Information Processing & Management*, 49(1):179–197.

L. Padró and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

R. Parker, D. Graff, J. Kong, K. Chen, and K. Maeda. 2011. English Gigaword Fifth Edition. LDC catalog ref. LDC2011T07.

J. Pustejovsky, P. Hanks, R. Saurí, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, et al. 2003. The TimeBank corpus. In *Corpus Linguistics*.

J. Pustejovsky, B. Ingria, R. Saurí, J. Castano, J. Littman, R. Gaizauskas, A. Setzer, G. Katz, and I. Mani. 2005. The specification language TimeML. *The Language of Time: A reader*, pages 545–557.

R. Saurí and T. Badia. 2012. Spanish TimeBank 1.0. LDC catalog ref. LDC2012T12.

R. Saurí, J. Littman, B. Knippen, R. Gaizauskas, A. Setzer, and J. Pustejovsky. 2006. TimeML Annotation Guidelines Version 1.2.1.

M. Taulé, M. A. Martı, and M. Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008)*.

N. UzZaman and J. Allen. 2010. TRIPS and TRIOS system for TempEval-2: Extracting temporal information from text. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 276–283. Association for Computational Linguistics.

N. UzZaman and J. Allen. 2011. Temporal Evaluation. In *Proceedings of The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

N. UzZaman, H. Llorens, and J. Allen. 2012a. Evaluating temporal information understanding with temporal question answering. In *Proceedings of IEEE International Conference on Semantic Computing*.

N. UzZaman, H. Llorens, J. F. Allen, L. Derczynski, M. Verhagen, and J. Pustejovsky. 2012b. TempEval-3: Evaluating Events, Time Expressions, and Temporal Relations. *CoRR*, abs/1206.5333.

N. UzZaman. 2012. *Interpreting the Temporal Aspects of Language*. Ph.D. thesis, University of Rochester, Rochester, NY.

M. Verhagen, R. Gaizauskas, F. Schilder, M. Hepple, J. Moszkowicz, and J. Pustejovsky. 2009. The TempEval challenge: identifying temporal relations in text. *Language Resources and Evaluation*, 43(2):161–179.

M. Verhagen, R. Saurí, T. Caselli, and J. Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62. Association for Computational Linguistics.