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

Outline

- Motivation
- TempEval-3
- TempEval-3 participants
- Summary and Future Work

Temporal Information Processing

In **1492**, Genoese explorer Christopher Columbus, under contract to the Spanish crown, **reached** several Caribbean islands, making first contact with the indigenous people. On **April 2, 1513**, Spanish conquistador Juan Ponce de León **landed** on what he called "La Florida"—the first documented European arrival on what would become the U.S. mainland. Spanish settlements in the region were followed by ones in the present-day southwestern United States that drew thousands through Mexico. French fur traders **established** outposts of New France around the Great Lakes; France eventually **claimed** much of the North American interior, down to the Gulf of Mexico. The first successful English settlements were the Virginia Colony in Jamestown in 1607 and the Pilgrims' Plymouth Colony in 1620. The 1628 chartering of the Massachusetts Bay Colony **resulted** in a wave of **migration**; by 1634, New England had been settled by some 10,000 Puritans. Between the late 1610s and the American Revolution, about 50,000 convicts were shipped to Britain's American colonies. Beginning in 1614, the Dutch settled along the lower Hudson River, including New Amsterdam on Manhattan Island.

Why is it important?

- Question Answering (QA)
- Summarization
- Visualization
- Natural Language Understanding

TempEval-3 vs earlier TempEval

- Size of corpus: 600K silver, 100K gold vs 50K
- End-to-end temporal relation processing task vs subtasks
- Full set of TimeML temporal relations vs reduced set used in earlier TempEvals
- Platinum test set
- Evaluation with temporal awareness score

Outline

- Motivation
- TempEval-3
- TempEval-3 participants
- Summary and Future Work

Data

- Reviewing Existing Corpora
  - TimeBank and AQUAINT
  - added missing entities and relations
  - converted all to full set of TimeML relations
- New Corpora
  - platinum corpus - annotated by the experts
  - silver corpus - merged with the SOA systems

Data

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

Task A: temporal expression

Temporal expression	Type	Value
DCT (given): March 1, 1998; 14:11 hours	TIME	1998-03-01T14:11:00
Sunday	DATE	1998-03-01
last week	DATE	1998-W08
mid afternoon	TIME	1998-03-01TAF
nearly two years	DURATION	P2Y
each month	SET	P1M

Table 1: Examples of normalized values and types for temporal expressions according to TimeML

Task B: events

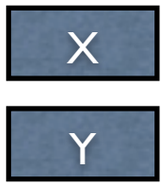
Event: something that happens or a dynamic property which holds the truth. Attributes: class, tense, aspect.

Class attribute:

1. Occurrence: die, crash, build
2. State: on board, alive
3. Reporting: say, report
4. I-Action: attempt, try, promise
5. I-State: believe, intend, want
6. Aspectual: begin, stop, continue
7. Perception: see, hear, watch, feel

Temporal Relations

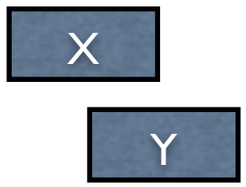
simultaneous
identity
during



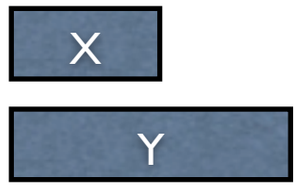
before /
after

iBefore /
iAfter

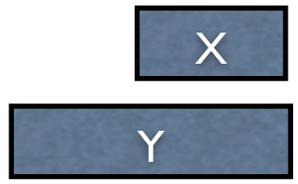
overlaps/
overlapped by



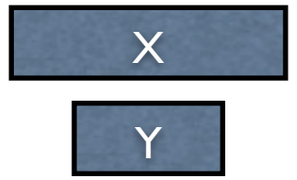
begins/
begun by



ends/
ended by



includes/
included by



Task ABC: temporal relations

- Task ABC - annotating temporal relations from raw text
- Task C - annotating relations given gold entities
- Task C relations only - annotating relations given gold entities and related pairs

Evaluation Metric - Task A and B

- Sorted with attribute F1 score - captures entity and attribute performance together

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

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

Evaluation Metric - Task ABC and C

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

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

where, G^+ is the closure of graph G and G^- is the reduced of graph G, where redundant relations are removed.

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

- F score -- temporal awareness score
- F score captures entity & relation performance
- F score can help to find overall best system

Outline

- Motivation
- TempEval-3
- TempEval-3 participants
- Summary and Future Work

Participants' Task A Systems

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 1: Automated approaches for TE3 Timex Extraction

Participants' Task A Performance

	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 1: Task A - Temporal Expression Performance.

Task A - Timex Observations

- Strategy: close competitions for all approaches
- Data: quality dataset helped more than larger dataset - silver data didn't help

Participants' Task B Systems

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 1: Automated approaches for Event Extraction

Participants' Task B Performance

	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 1: Task B - Event Extraction Performance.

Task B - Event Observations

- Strategy: machine learning systems dominant
- Data: larger dataset helped - silver data helped
- Linguistic Features: semantic features (lexical semantics and sentence-level semantic) helped, when executed properly

Task C: relation identification systems

Strategy	System	Training data	Classifier used	Linguistic Knowledge
Data-driven	ClearTK-1	TimeBank	SVM, Logit	*e-attr, ms*
	ClearTK-2	TimeBank + Bethard07	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 1: Automated approaches for TE3 TLINK Identification

Task C: relation classification systems

Strategy	System	Training data	Classifier used	Linguistic Knowledge
Data-driven	ClearTK-1	TimeBank	SVM, Logit	*ms, ls*
	ClearTK-2	TimeBank + Bethard07	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 1: Automated approaches for Relation Classification

Task C: Relation 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 1: Task ABC - Task C evaluation from raw text

	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 1: Task C - relation only: Relation Classification.

	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 1: Task C - TLINK Identification and Classification

Task C - Relations Observations

- Strategy: machine learning systems dominant
- Data: no one used silver data for final system
- Data: adding extra high quality data from external sources usually helped
- Features: using more linguistic features are important, but needs to be executed properly

TempEval-3 Spanish

- Two participants for temporal expressions
- One participant for event extraction
- TIPSemB-Freeling provided as SOA reference
- TE-3 participant HeidelTime outperformed in temporal expressions task

Outline

- Motivation
- TempEval-3
- TempEval-3 participants
- **Summary and Future Work**

Summary

- Larger corpus: 600K silver, 100K gold
- End-to-end temporal relation processing task
- Full set of TimeML temporal relations
- Platinum test set
- Evaluation with one score
- Found out what worked for each tasks

Future

- Release larger corpus for human review merging all participants' output
- Temporal QA can better evaluate temporal information understanding
  - Natural task; easy to create questions / annotate
  - Evaluates NLU; encourage for deep document level NLU
  - Can evaluate human vs system annotations
  - Can reason with any TimeML annotations

Acknowledgement

- 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.

Thank you!

Why Evaluate with Temporal QA

- Answering questions to judge understanding
- Creating questions easier than annotating
- Temporal IE performance might not reflect real task performance
- Completeness of human annotations
- Comparing human and system annotations

Temporal Question Taxonomy

- yes/no: Was Fein called after the killing?

- list: What happened after the crash?
- list between: What happened between the crash and today?
- factoid: When DT Inc. holders adopted a shareholder-rights plan?

during the crisis

Temporal QA System

- Take TimeML annotation as input
- Temporal reasoning with Timegraph (Miller and Schubert, 1990) - efficient
- Maintains necessary relations in graph to infer relations between all entities
- Can answer yes/no, list, factoid

Evaluating Temporal Information Understanding with Temporal Question Answering. Naushad UzZaman, Hector Llorens and James F. Allen. Proceedings of IEEE **International Conference on Semantic Computing**, Italy, September 2012.

Summary of Temporal QA

- Temporal QA can better evaluate temporal information understanding
  - Natural task; easy to create questions
  - Can evaluate human temporal annotations
- Developed a system for temporal QA
  - Can reason with any TimeML annotations
  - Released the toolkit

Metrics for Temporal Evaluation

Evaluation Metric	Recall	Precision
TempEval-2	$\frac{Ref \cap Sys}{Ref}$	$\frac{Sys \cap Ref}{Sys}$
Setzer et al.	$\frac{Ref^+ \cap Sys^+}{Ref^+}$	$\frac{Sys^+ \cap Ref^+}{Sys^+}$
Tannier and Muller	$\frac{Ref^- \cap Sys^-}{Ref^-}$	$\frac{Sys^- \cap Ref^-}{Sys^-}$
Our ACL'11 metric	$\frac{\lvert Ref^- \cap Sys^+ \rvert}{\lvert Ref^- \rvert}$	$\frac{\lvert Sys^- \cap Ref^+ \rvert}{\lvert Sys^- \rvert}$
Our Updated metric	$\frac{\lvert Ref^- \cap Sys^+ \rvert + w * \lvert (Sys^- - Ref^-) \cap Ref^+ \rvert}{\lvert Ref^- \rvert}$ [1]	$\frac{\lvert Sys^- \cap Ref^+ \rvert}{\lvert Sys^- \rvert}$

[1] where, $w = \frac{0.99}{(1+\lvert \#Ref^+ \rvert - \lvert Ref^- \cap Sys^+ \rvert)}$

Improvement

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

* G^- = Reduced Graph of G, which includes only the core relations

$$Recall = \frac{|Ref^- \cap Sys^+| + w * |(Sys^- - Ref^-) \cap (Ref^+ - Ref^-)|}{|Ref^-|}$$

* $(Ref^- \cap Sys^+)$ captures system explicit relations that are found in gold explicit relations

* $(Sys^- - Ref^-)$ captures implicit system relations in terms of Ref^-

* $(Ref^+ - Ref^-)$ captures implicit gold relations

* $(Sys^- - Ref^-) \cap (Ref^+ - Ref^-)$ captures implicit system relations that exists in gold implicit relations