The importance of a semantics for semantic annotation: the temporal case

Philippe Muller
Alpage Project-Team & IRIT
INRIA & Toulouse University
muller@irit.fr

ILIKS meeting
Aix, June 2011

Joint work with Pascal Denis (Alpage),
draws from papers at Coling 2010 and Ijcai 2011

Overview

- objective: temporal annotation of relations between events in texts
- consensual semantics in NLP: relations between time intervals
- importance of reasoning over (relational) representations
  - for human annotation: clear instructions
  - for predictions by system: control of coherence
  - for comparison/evaluation of human annotations, system predictions
  - for translating between different representation schemes
- here:
  - comparing different representations schemes wrt predictions
  - using reasoning to improve automatic prediction

Temporal ordering

Important task for language understanding consists in recovering "chronology" of temporal entities described in texts

President Joseph Estrada on Tuesday $_{t_4}$ condemned $_{e_1}$ the bombings $_{e_5}$ of the U.S. embassies in Kenya and Tanzania and offered $_{e_{12}}$ condolences to the victims. [...] In all, the bombings $_{e_{10}}$ last week $_{t_5}$ claimed $_{e_4}$ at least 217 lives.

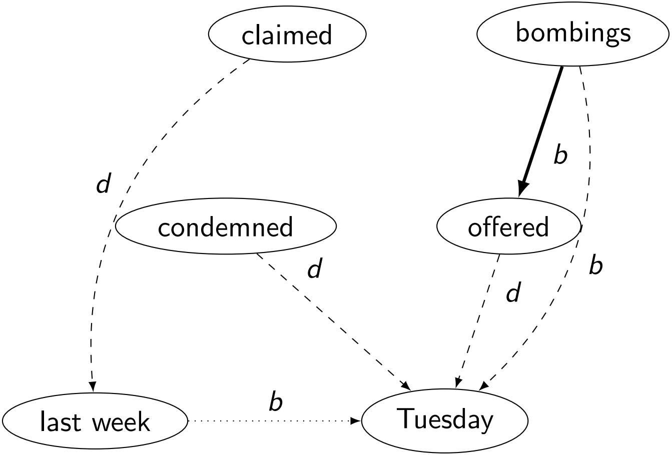
Ordering : e_5 before e_1, e_1 during t_4, e_4 before e_1, ...
Relation types : time-time, event-time, event-event

ISO specification: ISO-TimeML within ISO TC 37/SC 4 (TLINKS)

Importance of temporal reasoning

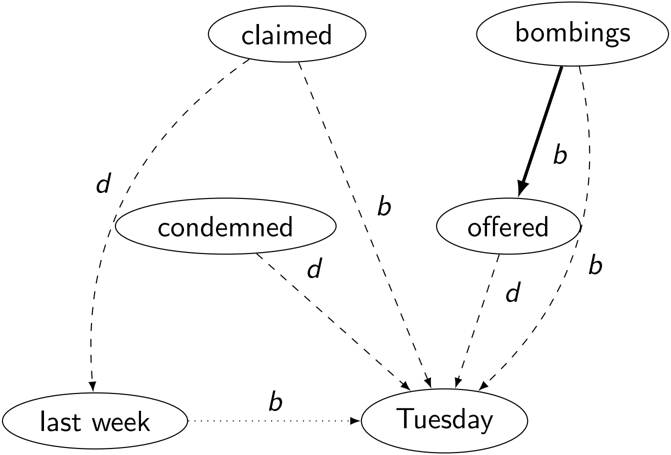
- Temporal relations have logical properties associated with them: e.g.,
  - $before(e_1, e_2) \models after(e_2, e_1)$ (symmetry of precedence)
  - $before(e_1, e_2), before(e_2, e_3) \models before(e_1, e_3)$ (transitivity of precedence)
  - $before(e_1, e_2), during(e_3, e_2) \models before(e_1, e_3)$ (transitivity of precedence + inclusion)
- These properties are important because:
  - They restrict the coherent graphs that can be built for a set of events: e.g., $before(e_1, e_2), during(e_3, e_2), after(e_1, e_3)$
  - They allow us to compare different descriptions of the same situation: $before(e_1, e_2), during(e_3, e_2) \equiv$ $before(e_1, e_2), during(e_3, e_2), before(e_1, e_3)$

d=during, b=before



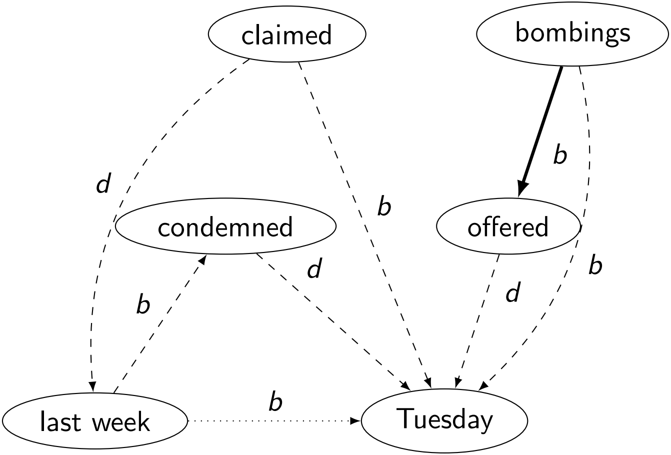
Temporal ordering: graph representation
with inferences

d=during, b=before

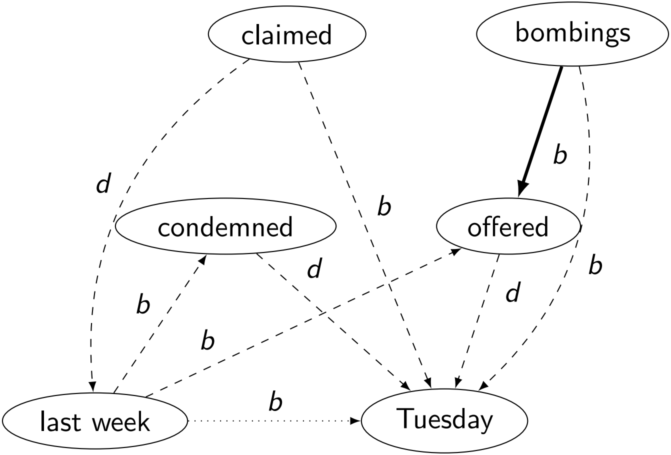


Temporal ordering: graph representation
with inferences

d=during, b=before

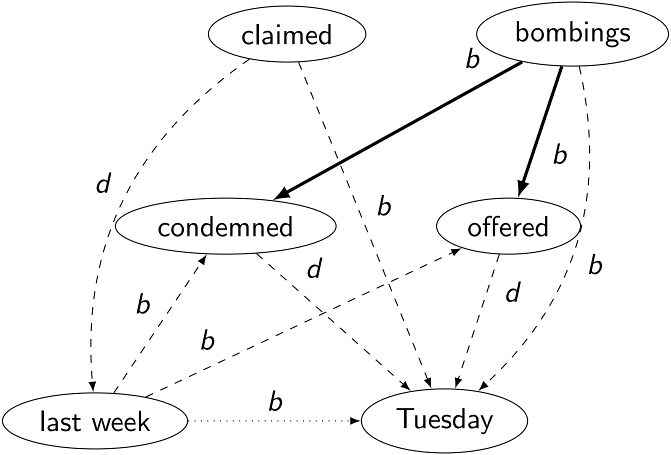


d=during, b=before



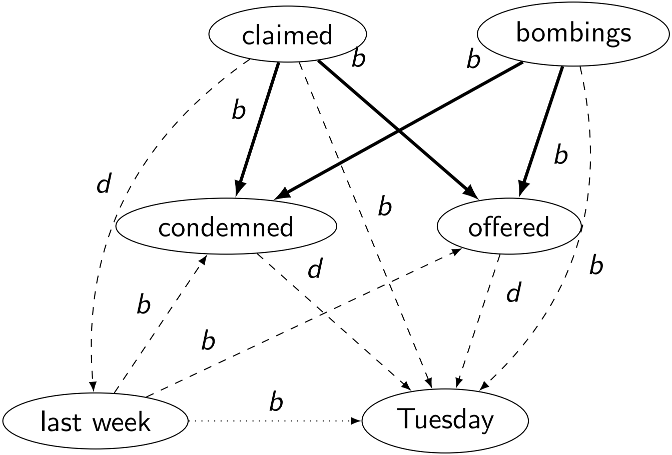
Temporal ordering: graph representation
with inferences

d=during, b=before

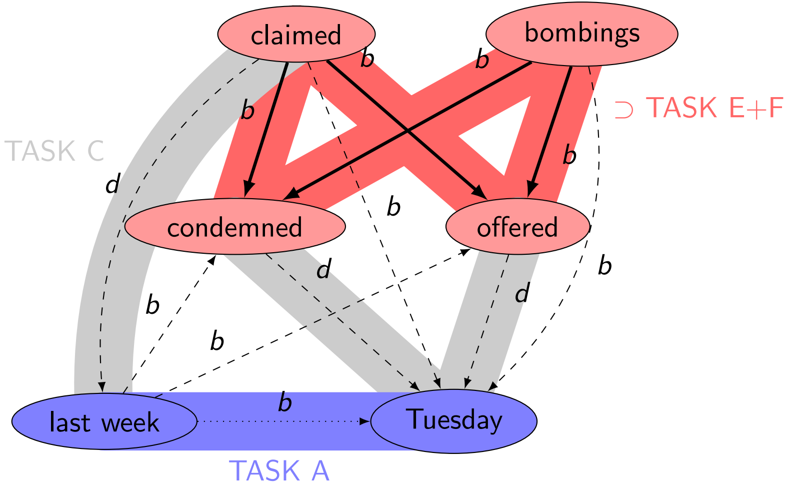


Temporal ordering: graph representation
with inferences

d=during, b=before



d=during, b=before



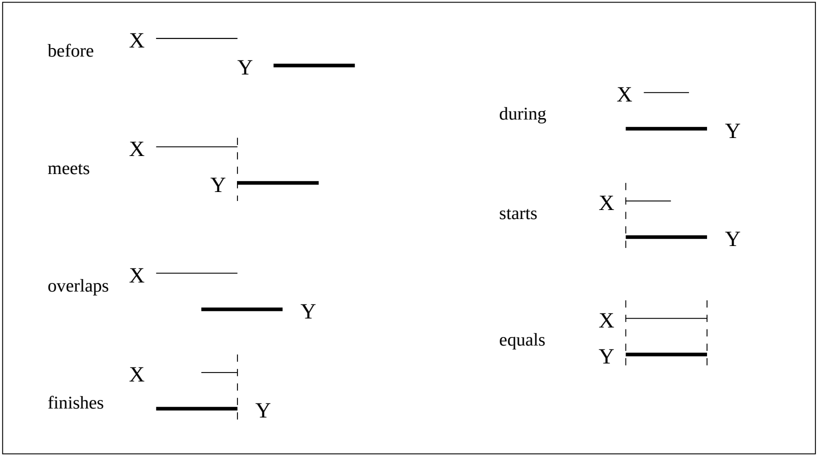
Annotation choices versus knowledge representation ?

- all possible orderings between time intervals w.r.t. to endpoint ordering: Allen relations [Allen, 1983]
- ISO TimeML specification : almost the same, excludes partial overlaps (TimeBank)
- TempEval campaign: much vaguer relations, supposedly easier to annotate
- other choices are possible: e.g. endpoint, semi-intervals, Bruce's 7 relations
- balance between feasability, naturalness of annotation and power of representation ?

→ separate annotation from reasoning

→ necessity of conversions between levels of representations

Interval-interval relations



Allen's thirteen relations between two temporal intervals

Correspondances

<table>
<tr><th>TimeML</th><th>Allen</th><th>Bruce</th><th>Tempeval</th></tr>
<tr><td>BEFORE
IBEFORE</td><td>before
meet</td><td>before</td><td>before</td></tr>
<tr><td>(absent)</td><td>overlaps</td><td>overlaps</td><td rowspan="4">overlaps</td></tr>
<tr><td>STARTS
IS_INCLUDED
FINISHES</td><td>starts
during
finishes</td><td>included</td></tr>
<tr><td>(absent)</td><td>overlapsi</td><td>is-overlapped</td></tr>
<tr><td>IS_STARTED
INCLUDES
IS_FINISHED</td><td>startsi
duringi
finishesi</td><td>includes</td></tr>
<tr><td>IAFTER
AFTER</td><td>meeti
beforei</td><td>after</td><td>after</td></tr>
<tr><td>SIMULTANEOUS</td><td>equals</td><td>equals</td><td>equals</td></tr>
</table>

A relation ranging over multiple cells is equivalent to a disjunction of all the relations within these cells.

What kind of reasoning ?

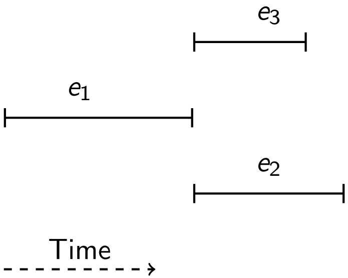
- semantic enrichment, eg for extraction, machine learning (deductions)
- comparison of different annotations (equivalences)
- control of coherence
- but: need to be computationally feasible
- $\rightarrow$ restricted forms of temporal reasoning: constraint languages, aka relational algebras

Algebras of relations

- a set of base relations, jointly exhaustive and mutually exclusive
- any relation between domain objects is a disjunction between base relations, considered as a set of relations
- any two relations can be composed to yield a new relation:
  $\mathit{before}(e_1, e_2), \mathit{during}(e_3, e_2) \rightarrow \mathit{before}(e_1, e_3)$
  noted : $\text{before} \circ \text{during} = \text{before}$
- in general, composition is disjunction of compositions between base relations
- this defines an algebra on relations with operations $\cup, \cap, \circ$
- composition of relations is guaranteed to reach a fixed point
  $\rightarrow$ saturation used for comparison of two annotations
- or signal an inconsistency $\rightarrow$ coherence control

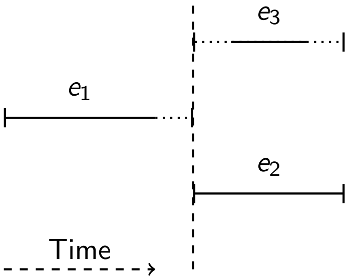
Allen, Bruce and Tempeval relations all define algebras

Importance of inferential power



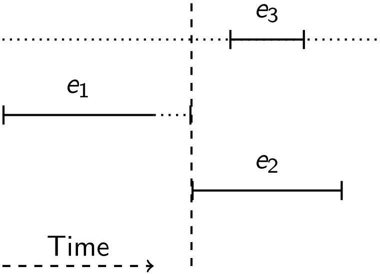
Allen: $(e_1 \text{ meet}_a\ e_2 \wedge e_3 \text{ starts}_a\ e_2) \rightarrow e_1 \text{ meet}_a\ e_3$

Importance of inferential power



Bruce: $(e_1 \text{ before}_b\ e_2 \wedge e_3 \text{ during}_b\ e_2) \rightarrow e_1 \text{ before}_b\ e_3$

Importance of inferential power



Tempeval: $(e_1 \text{ before}_t\ e_2 \wedge e_3 \text{ overlaps}_t\ e_2) \rightarrow e_1 \{\text{before}_t, \text{overlaps}_t\} e_3$

Experiment 1: comparing algebras for learning
Coling 2010

- how can temporal reasoning be best used to learn temporal orderings ?
- $\rightarrow$ compare the impact of using different temporal relation sets
- in particular, what is the best trade-off between:
  - how easy it is to learn a given relation set
    $\approx$ number x generalizations captured
  - how much new information can be inferred by the representations produced by each relation set
    $=$ "inferential power"
  - how accurate and coherent are the predicted complete temporal orderings?

Methodology

- Use OTC = TimeBank + ACQUAINT corpus
- Learn event-pair classifiers based on the different algebras (base relations only): Allen, Bruce, TempEval
- Evaluate algebra specific models on two tasks
    1. classification task: event pairs annotated with a temporal relation
    2. "global" task of producing complete (i.e., closed) event-event graphs (coherence enforced)
- For each algebra in which we learn and predict, we can evaluate in all algebra that are vaguer

Testing on whole graphs

As most approaches:

only label event-pairs given by the gold annotation

Two types of greedy decoding:

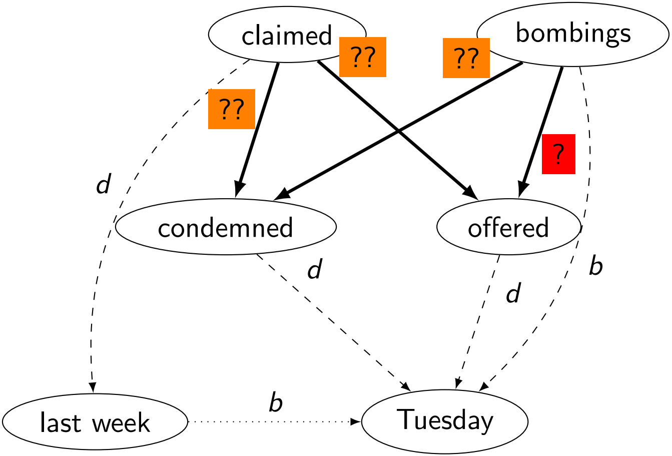
1. "argmax" decoding: pick relation with highest probability for each event pair (no coherence check)
2. "natural reading order" decoding: pick most probable relation that preserves global coherence (this implies saturation after classification)

Types of evaluation

1. Classification accuracy on annotated relations (no coherence check)
2. Precision / Recall on closed graphs
   - "Strict" measures:
     only compare the sets of simple temporal relations
   - "Relaxed" measures:
     compare the overlaps of sets of relation disjunctions
     (universal disjunction everywhere get .33!)
   - inconsistent graphs intepreted as making no prediction
     (only recall is penalized)
   - predictions in one algebra can be converted and evaluated into any another one for relaxed measures

Input to the task

T-T and E-T relations given (easier tasks)
gold annotation event-pairs
event-pairs from saturated gold annotation graphs



Event-Event TLINK classification accuracy

target : gold annotation event-pairs

	Allen	TempEval
Allen	47.0	48.9
Bruce	N/A	49.3
TempEval	N/A	**54.0**

- As expected, algebra-specific classifiers have the best accuracy when evaluated in their own algebra
- Best absolute accuracy performances are given by the vaguer TempEval-classifier

targets : saturated gold annotation graphs

		Relaxed F1	Strict F1
argmax	Allen	**51.5**	**52.7**
	Bruce	42.1	25.9
	Tempeval	36.5	21.2
NRO	Allen	**51.3**	**59.9**
	Bruce	49.5	21.2
	Tempeval	36.5	21.2

- not very interesting for "strict" metrics: Bruce and Tempeval can only predict "equals"
- Allen-based system strongly outperforms the systems using models trained on vaguer algebras due to the largely under-specified representations these produce

targets : saturated gold annotation graphs

		relaxed F1	strict F1
argmax	Allen	55.5	53.6
	Bruce	**57.3**	**53.8**
	Tempeval	48.2	29.1
NRO	Allen	65.3	51.8
	Bruce	**68.5**	**52.9**
	Tempeval	48.2	29.1

- Allen- and Bruce-based systems significantly outperform TempEval system, on its evaluation home ground and even tough their classification accuracy was lower
- Bruce-based system performs best, providing the best trade-off between “learnability” and expressive power (not by much)

Experiment 2: predict coherent structures
Ijcai 2011

- goal: predict consistent temporal structures
- learn local relational model (classic)
- use reasoning during decoding to produce best globally coherent set of relations
  - classic: use Integer Linear Programming (ILP) translation to enforce coherence
  - new: use the full set of relations
- new: translates annotations as end point representations to make it computationally practical
- new: doing it without taking all the reference pairs as given
  - on reference self-connected temporal subgraphs
  - on heuristically determined meaningful subgraphs

Using points instead of intervals

$R \in Allen,\ r_i \in \{\prec, \succ, =\}$

$$R(I_1, I_2) \equiv \quad r_1(I_1^-, I_2^-) \ \wedge \ r_2(I_1^+, I_2^-) \ \wedge \ r_3(I_1^-, I_2^+) \ \wedge \ r_4(I_1^+, I_2^+)$$

Composition

$\circ$	$\prec$	$\preceq$	$\succ$	$\succeq$	$=$
$\prec$	$\prec$	$\prec$			$\prec$
$\preceq$	$\prec$	$\preceq$			$\preceq$
$\succ$			$\succ$	$\succ$	$\succ$
$\succeq$			$\succ$	$\succeq$	$\succeq$
$=$	$\prec$	$\preceq$	$\succ$	$\succeq$	$=$

Conversion end-points/interval

Allen	order/endpoints
b	$(\prec, \prec, \prec, \prec)$
m	$(\prec, =, \prec, \prec)$
o	$(\prec, \succ, \prec, \prec)$
s	$(=, \succ, \prec, \prec)$
d	$(\succ, \succ, \prec, \prec)$
f	$(\succ, \succ, \prec, =)$

Results on reference pairs
the "easy" task

System	baseline zero	baseline before	nro	ilp
Recall	26.01	37.93	20.08	49.80

- TimeBank only (harder than OTC)
- strict evaluation
- zero = do nothing but assumes perfect E-T and T-T relations
- before = order events in order of text
- without coherence enforcement, local classifier yields 82% inconsistent graphs
- recall is (somewhat improperly) called accuracy in comparable studies

Results without assuming reference pairs
the “real” task

wrt connected components

System	Precision	Recall	F1-score	Inco.
ilp	33.02	54.07	41.00	5.93
nro	49.98	17.02	25.40	0.00
before	6.22	37.93	10.69	0.00

NB

- local classifiers :
  88% inconsistent graphs on connected components,
- tested also on heuristic subgraphs: not very good so far

Take home message(s)

preaching to the choir

- semantic annotations need precise semantics
- semantic annotations need a deduction model
- different schemas can co-exist but should be related within a formal representation framework
- **deduction is good for you**
  (for evaluation, comparison, prediction)

Perspectives

- enriched representations $\rightarrow$ a lot of evaluation issues
  (cf joint work with Xavier Tannier, JAIR 2011)
- similar relational problems:
  - spatial relations, although inference seems less productive
  - discourse relations: although semantics constraints and equivalences not well established
    (work in progress in Alpage team: Charlotte Roze)
- integration within human annotation process ?
  (cf work of Mark Verhagen at Brandeis)

References

Pascal Denis and Philippe Muller.
Comparison of different algebras for inducing the temporal structure of texts.
In *Proceedings of Coling 2010*, pages 250–258, Beijing, 2010.

Pascal Denis and Philippe Muller.
Predicting globally-coherent temporal structures from texts via endpoint inference and graph decomposition.
In *Proceedings of IJCAI 2011*, pages xx–xx, 2011.

Xavier Tannier and Philippe Muller.
Evaluating temporal graphs built from texts via transitive reduction.
Journal of Artificial Intelligence Research, (40):375–413, 2011.