

Utilizing Temporal Information in Topic Detection and Tracking

Juha Makkonen and Helena Ahonen–Myka

`{jamakkon,hahonen}@cs.helsinki.fi`

University of Helsinki – Department of Computer Science

Outline

- Introduction
- Topic Detection and Tracking
- Resolving temporal expressions
 - Recognition
 - Formalization
 - Comparison
- Experiments
- Future Work

Introduction

- Temporal expressions are often omitted.
 - their extraction requires tools,
 - they have to be formalized in order to be of any use,
 - comparing formalizations is sometimes tricky.
- By no means a novel idea
 - in AI to form chronologies of events,
 - in question answering to extract a fact,
 - in databases, diagnosing systems, dialog systems ...
- We want to measure the temporal *similarity* of two documents.

Topic Detection and Tracking

- TDT system monitors news broadcasts in order to
 - *detect* new, previously unreported events, and to
 - *track* the development of the detected events.
- The focus is on *news events*: something untrivial taking place at a specific time and place.
- A *topic* is understood as as is an event or an activity, along with all related events and activities.
- The news stream that is monitored is intrinsically sensitive to time.

Resolving Temporal Expressions

- An expression can be
 - explicit: “*the 19th of August 2003*”,
 - implicit: “*today*”, “*Tuesday afternoon*”, or
 - vague: “*since April*”, “*a couple of weeks ago*” .
- The evaluation is based on a point of reference. “*The winter of 1974 was cold. The next winter will be colder.*”
“*The winter of 1974 was cold. The next winter was colder.*”
- Resolving the meaning of the latter winter requires
 - the *reference time* or the *utterance time* and
 - the tense of the relevant verb.

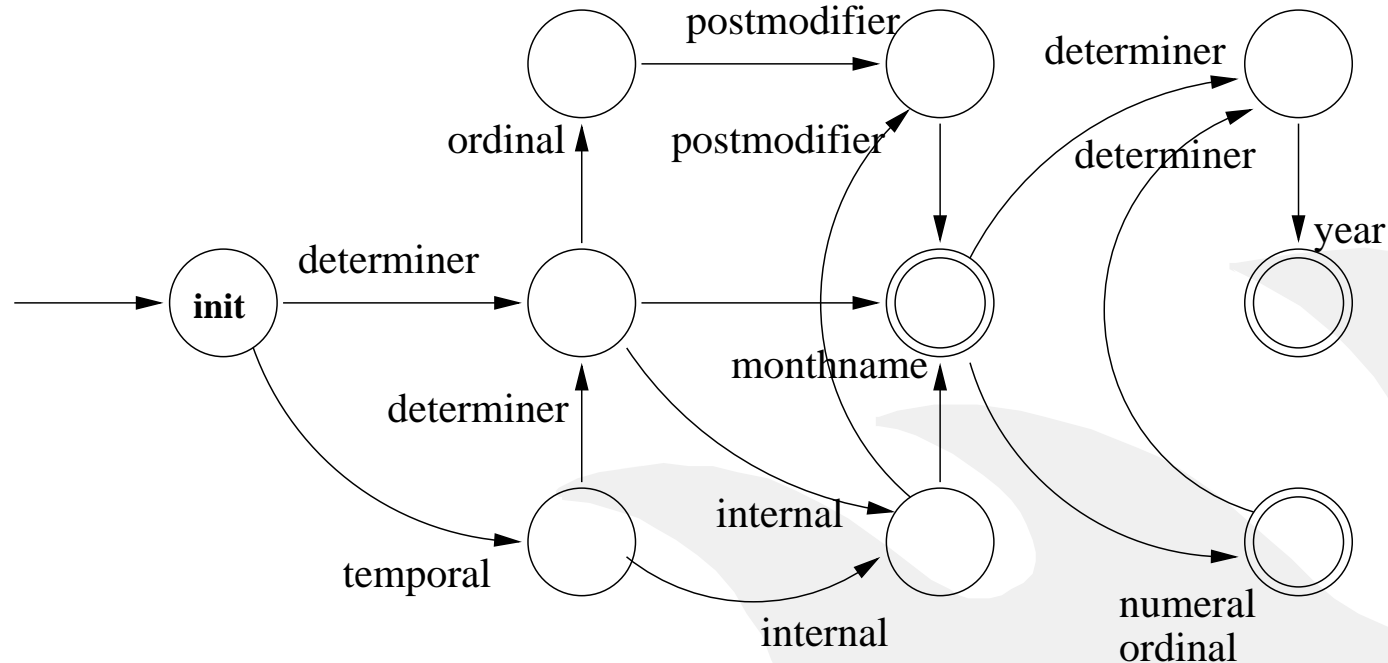
Recognition

- The relevant terms are split into categories.

<i>category</i>	<i>terms</i>
baseterm	day, week, weekday, month, monthname, quarter, season, year, decade
indexical	yesterday, today, tomorrow
internal	beginning, end, early, late, middle
determiner	this, last, next, previous, the
temporal	in, on, by, during, after, until, since, before, later
postmodifier	of, to
numeral	one, two, ...
ordinal	first, second, ...
adverb	ago
meta	throughout
vague	some, few, several
recurrence	every, per
source	from

Recognition

- The categories are used to build automata.



*“The strike started **on the 15th of May 1919**. It lasted **until the end of June**, although there was still turmoil in **late January next year**”.*

Formalization

- We map the expressions onto a *calendar*
 - a time-line – points with precedence relation,
 - a set of granularities (year, month, week, ...)
note: March, Thursday and weekend are also granularities.
 - a set of conversion functions between granularities.
- The expressions are mapped as intervals $[t_{start}, t_{end}]$ of the bottom granularity which in our case is *day*.

Formalization

- The baseterm of the expression defines interval.
- The non-baseterms are interpreted as shift and span functions that modify the start and end points.
 - shift: this, next, last, 3 weeks ago, etc.
 - span: until, before, after, from, etc.
- the length of the interval is modified by internals
 - in the beginning of 1970s, late May, etc.

Comparison

- We want to measure the temporal similarity of two documents, i.e., how much the references overlap.
- When comparing the intervals of two documents
 - compare pairwise all intervals
 - $\text{similarity} = 2 * \text{overlap} / \text{size of the intervals}$
 - take the average of the best matches for each interval.
- The outcome measures how well the references of one document cover those of the other.

Experiments

- Data: transcribed TV and radio broadcasts and online news.
 - 8595 documents from the TDT2 corpus.
 - 2383 documents were labeled to one of 35 events.
- Temporal expression recognition with 1417 sentences

<i>type</i>	<i>freq</i>	<i>recognition</i>	<i>canonization</i>
simple	326	0.98	0.93
composite	209	0.85	0.66

- Verbs like *to schedule* , *to plan* or *to expect* gave hard time.
- “*The meeting was scheduled for Monday.*” Which one?

Experiments

- The distribution of temporal relations

<i>relation</i>	<i>same event</i>	
	<i>yes</i>	<i>no</i>
before	0.761	0.831
meets	0.001	0.000
overlaps	0.016	0.008
begins	0.010	0.006
falls within	0.168	0.122
finishes	0.010	0.008
exact	0.072	0.056

Experiments

- Temporal similarity is higher when documents are relevant.

<i>average of</i>	<i>same event</i>	<i>different event</i>	<i>ratio of yes/no</i>
sum of pairwise	0.0034	0.0023	1.4783
max of pairwise	0.0059	0.0040	1.4750

- Finding the best-match for each interval does not pay off.
- A better accuracy on formalization would help.
- What is the meaning of “*three years ago?*”
- How to represent informativeness?

Future Work

- Improvement of the composite expression processing
 - more work on the automata
- Introduction of vagueness:
 - an expression would be formalized as probability distributions on the timeline
 - similarity could be Kullback-Leibler, for instance.
- Survey of the behaviour of the temporal expressions
 - how the references distribute per medium?
 - the first story compared to the following ones?

The End

Thank you

