Information Extraction: Temporal expression identification and normalization

Vivi Nastase

Summer semester 2012, ICL, University of Heidelberg
The importance of temporal expressions

Q: Is Bill Clinton currently the president of the United States?
The importance of temporal expressions

Q: Is Bill Clinton currently the president of the United States?

↓

April 2012
The importance of temporal expressions

Q: Is Bill Clinton currently the president of the United States?

April 2012

Q: Is Bill Clinton the president of the United States in April 2012?
Q: Who is the president of the United States in April 2012?
In 1957, Tolkien was to travel to the United States to accept honorary degrees ... . He retired two years later from his professorship at Oxford.

“The Adventures of Tom Bombadil” was published in 1962, three years after Tolkien retired from his professorship at Oxford.

... Tolkien makes a brief allusion to the future of Middle-earth in a letter written in 1958. The following year, after his retirement from teaching at Oxford, he ...
The importance of temporal expressions

Q: When did J.R.R. Tolkien retire from his professorship at Oxford?

In **1957**, Tolkien was to travel to the United States to accept honorary degrees ... . He **retired** two years later from his professorship at Oxford.

“The Adventures of Tom Bombadil” was published in **1962**, three years after Tolkien **retired** from his professorship at Oxford.

... Tolkien makes a brief allusion to the future of Middle-earth in a letter written in **1958**. The following year, after his **retirement** from teaching at Oxford, he ...
The importance of temporal expressions

Q: When did J.R.R. Tolkien retire from his professorship at Oxford?

In **1957**, Tolkien was to travel to the United States to accept honorary degrees ... . He **retired two years later** from his professorship at Oxford.

“**The Adventures of Tom Bombadil**” was published in **1962**, **three years after** Tolkien **retired** from his professorship at Oxford.

... Tolkien makes a brief allusion to the future of Middle-earth in a letter written in **1958**. The **following year**, after his **retirement** from teaching at Oxford, he ...
The importance of temporal expressions

Q: When did J.R.R. Tolkien retire from his professorship at Oxford?

In **1957**, Tolkien was to travel to the United States to accept honorary degrees ... He **retired two years later** from his professorship at Oxford.

“The Adventures of Tom Bombadil” was published in **1962**, **three years after** Tolkien **retired** from his professorship at Oxford.

... Tolkien makes a brief allusion to the future of Middle-earth in a letter written in **1958**. The **following year**, after his **retirement** from teaching at Oxford, he ...

---

1957 + 2

1962 - 3

1958 + 1
Issues related to temporal expressions

- Recognizing temporal expressions
  - absolute April 30\textsuperscript{th}, 2012; spring of 2012
  - relative today, yesterday, last year
  - durations two hours, one second
  - mixed? two months last year

- Linking temporal expressions to events

- Normalizing time expressions and reasoning about time
  - what is the basic temporal unit
  - representing temporal meaning
Recognizing temporal expressions

Hint: TEs are phrases with temporal **lexical triggers** as their heads.

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>noun</td>
<td>morning, noon, night, dusk, dawn, ...</td>
</tr>
<tr>
<td>proper noun</td>
<td>April, Monday, Easter, Labour Day, ...</td>
</tr>
<tr>
<td>adjective</td>
<td>recent, past, current, annual, ...</td>
</tr>
<tr>
<td>adverb</td>
<td>hourly, daily, weekly, monthly, yearly, ...</td>
</tr>
</tbody>
</table>
Issues

• ambiguity:
  Sunday Bloody Sunday is noted for its militaristic drumbeat, harsh guitar, and melodic harmonies. Among the seminal texts of the 20th century, 1984 is a rare work that grows more haunting as its futuristic purgatory becomes more real.
  USA Today, 20th Century Fox, Daily Telegraph

• variety in length:
  The IE course is scheduled on Mondays.
  I traveled for the whole Monday night.

• anaphoric expressions:
  Evelyn has seen 80 winters. This, she says, was the coldest.
Recognizing temporal expressions

What to do?

- Identify the fragment that expresses temporal information (segmentation)
- Identify the type of time expression (absolute / relative)

How to do it?

bootstrapping based on seed examples and patterns

rule-based using partial parsing or chunking

statistical sequence classifiers based on standard token-by-token IOB (Inside-Outside-Begin) encoding

learning based on annotated examples – constituent-based classification
Recognizing temporal expressions

- **bootstrapping** based on seed examples and patterns
  - **rule-based** using partial parsing or chunking
  - **statistical** sequence classifiers based on standard token-by-token IOB (Inside-Outside-Begin) encoding
  - **learning** based on annotated examples – constituent-based classification
Bootstrapping in general

Seokhwan Kim et al., 2011: Semi-supervised Information Extraction
Bootstrapping in general

Start either with a non-empty set $S$ of seed examples or a non-empty set $P$ of patterns (let’s assume examples):

1. find all occurrences of the examples in $S$ in your text collection
2. extract [and rank] patterns surrounding the examples
3. add the [highest ranking] extracted patterns to $P$
4. use the patterns in $P$ to find additional examples
5. add the extracted examples to $S$, go to step 1
Bootstrapping for temporal expressions

Poveda et al. 2009 An analysis of bootstrapping for the recognition of temporal expressions
Semantic classes:

- automatically generated word clusters (Lin, 1998)
- manually assembled word lists:
  - cardinals (1, 3, ...); ordinals (1\textsuperscript{st}, 30\textsuperscript{th}, ...)
  - days (Monday, today, ...); months (January, ...)
  - date trigger words (day, week, ...)
  - time trigger words (hour, minute, ...)
  - frequency adverbs (hourly, monthly, ...)
  - date and time adjectives (two-day, week-long, ..., three-hour, minute-long, ...)
Bootstrapping for temporal expressions

Grammar for patterns:

\[
\text{pattern} ::= \text{prefix SEP infix SEP postfix SEP (modifiers)}^* \\
\text{prefix} ::= (\text{pattern-elem})^* \\
\text{infix} ::= (\text{pattern-elem})^+ \\
\text{postfix} ::= (\text{pattern-elem})^* \\
\text{pattern-elem} ::= \text{FORM (token-form)} | \\
\text{SEMCLASS (token-form)} | \\
\text{POS (pos-tag)} | \\
\text{LEMMA (lemma-form)} | \\
\text{SYN (syn-type, head)} | \\
\text{SYN-SEM (syn-type, head)} \\
\text{modifiers} ::= \text{COMPLETE-PHRASE}
\]

*end of December* $\rightarrow$ LEMMA(end) LEMMA(of) SEMCLASS(MONTH)

*end of January 2009* $\rightarrow$ LEMMA(end) LEMMA(of) SEMCLASS(MONTH) COMPLETE-PHRASE
Score and filter patterns

- \( \mathcal{E} \) – current set of instances \( e_i \)
- \( \mathcal{I}_p \) – set of instances of pattern \( p \)

\[
\text{freq}_{-}\text{sc}(p) = |\mathcal{I}_p \cap \mathcal{E}| - \text{coverage of a pattern}
\]

\[
\text{prec}_{-}\text{sc}(p) = \frac{\text{freq}_{-}\text{sc}(p)}{|\mathcal{I}_p|} = \frac{|\mathcal{I}_p \cap \mathcal{E}|}{|\mathcal{I}_p|}
\]
Score and filter new instances

$\mathcal{E}$ – current set of instances $e_i$
$\mathcal{C}_e$ – set of contexts of infix of $e$

$$sc(e) = \lambda_1 sim\_sc(e) + \lambda_2 pc\_sc(e) + \lambda_3 ctxt\_sc(e)$$

$sim\_sc(e)$ similarity score:

$$sim\_sc(e) = \frac{\sum_{i=1}^{n} \log(1 + Sim(w_i))}{n}$$

$$Sim(w_i) = \sum_{j=1}^{||\mathcal{E}||} \max(sim(w_i, w_{e_j,1}), \ldots, sim(w_i, w_{e_j,|e_j|}))$$

$pc\_sc(e)$ phrase completeness score $= \frac{c(INFIX)}{c(*INFIX*)}$

c$ctxt\_sc(e)$ context based score $= \frac{c(mfw, \mathcal{C}_e)}{c(mfw)}$

$mfw$ – most frequent word in $\mathcal{C}_e$
Particularities of bootstrapping for temporal expressions

- syntactic information
- distributional semantics
- pattern subsumption analysis
- variable length patterns
Recognizing temporal expressions

**bootstrapping** based on seed examples and patterns

**rule-based** using partial parsing or chunking

**statistical** sequence classifiers based on standard token-by-token
  IOB (Inside-Outside-Begin) encoding

**learning** based on annotated examples – constituent-based classification
Rule-based temporal expression recognition

Negri and Marseglia, 2004 *Recognition and normalization of temporal expressions*

Mazur and Dale, 2007 *A rule based approach to temporal expression tagging*

- hand-crafted rules (≈ 1500 in Chronos)
- detect temporal expressions based on lexical triggers
- delimit the relevant context (bracketing) surrounding the lexical triggers
  *beginning, end, previous, next, ago, later, ...*
Basic rules

The early 1990s

<table>
<thead>
<tr>
<th>PATTERN</th>
<th>t1 t2 t3</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>[pos = &quot;DT&quot;]</td>
</tr>
<tr>
<td>t2</td>
<td>[lemma = &quot;early&quot;]</td>
</tr>
<tr>
<td>t3</td>
<td>[pred = decade-p]</td>
</tr>
</tbody>
</table>

**OUTPUT**

```xml
<TIMEX2 val="?"
  type="T-ABS"
  mod="START">
  t1 t2 t3
</TIMEX2>
```
Basic rules

Consider *triggers + context* to fill in (TIMEX2 attributes):

- **MOD**  *more than, approximately ...*
- **SET**  *every, twice a ...*
- **DIR**  *before, ago, during ...*
Basic rules

Consider *triggers + context* to fill in (TIMEX2 attributes):

- **MOD** *more than, approximately* ...
- **SET** *every, twice a* ...
- **DIR** *before, ago, during* ...

and (Temporary attributes):

- **type** *absolute / relative*
- **cat** *second / minute / hour / day, ...*
- **op** = / + / -
- **quant** $\geq 0$
Composition rules

... the whole Monday night ...

... the whole Monday ... / ... Monday night ... / ... the whole Monday night ...

<table>
<thead>
<tr>
<th>PATTERN</th>
<th>T-EXP1</th>
<th>T-EXP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-EXP1</td>
<td>[start = n]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[end = m]</td>
<td></td>
</tr>
<tr>
<td>T-EXP2</td>
<td>[start = o → n ≤ o m]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[end = p → o p ≤ m]</td>
<td></td>
</tr>
</tbody>
</table>

OUTPUT

| T-EXP1 | [start = n]|
|        | [end = m]  |
Recognizing temporal expressions

- **bootstrapping** based on seed examples and patterns
- **rule-based** using partial parsing or chunking
- **statistical** sequence classifiers based on standard token-by-token IOB (Inside-Outside-Begin) encoding
- **learning** based on annotated examples – constituent-based classification
Sequence labeling as classification

Classify an element of a sequence as B (begin), I (inside), O (outside) the chunks of interest.

Jurafsky and Martin, 2009 *Speech and text processing*
### Sequence labeling temporal expressions

**Commonly used features:**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Token</td>
<td>the target token to be labeled</td>
</tr>
<tr>
<td>POS</td>
<td>part of speech of the target token</td>
</tr>
<tr>
<td>Tokens in window</td>
<td>bag of tokens in the window around the target</td>
</tr>
<tr>
<td>POS in window</td>
<td>bag of POS in the window around the target</td>
</tr>
<tr>
<td>Chunk tags</td>
<td>base-phrase chunk tag for target and words in the window</td>
</tr>
<tr>
<td>Lexical triggers</td>
<td>presence in a list of temporal terms</td>
</tr>
</tbody>
</table>
Sequence labeling with HMMs

Maximize $P(X|O, \lambda)$

$X = x_1...x_n$ – sequence of hidden variable values

$O = o_1...o_n$ – observations

$\lambda = (A, B)$

$\lambda :$

$A$

$a_{ij} = p(x_i|x_j)$ transition probabilities

$a_{0j} = p(x_j)$ initial state probabilities

$B$

$b_i(o_k) = p(o_k|x_i)$ emission probabilities
Sequence labeling with HMMs

The Viterbi Algorithm

\[ v_1(j) = a_{0j}b_j(o_1) j = 1, N \]

\[ v_t(j) = \max_i v_{t-1}(i) a_{ij}b_j(o_t) j = 1, N \]

\[ \text{back}(j) = \arg \max_i v_{t-1}(i) a_{ij}b_j(o_t) \]
IOB sequence labeling with HMMs

- tokenize text
- split text into sentences (our sequences)
- hidden variable possible values: $I, O, B$
- estimate $\lambda$ from an annotated corpus

$$a_{ij} = p(x_i|x_j) = \frac{c(x_i, x_j)}{c(x_j)}$$

$$b_j(o_t) = \frac{c(x_j, o_t)}{c(x_j)}$$
Recognizing temporal expressions

**bootstrapping** based on seed examples and patterns

**rule-based** using partial parsing or chunking

**statistical** sequence classifiers based on standard token-by-token
    IOB (Inside-Outside-Begin) encoding

**learning** based on annotated examples – constituent-based classification
Constituent-based recognition of temporal expressions

- segmentation – based on syntactic phrases
- supervised classification (TE /not TE)
- features similar to those used in sequence labeling
Normalizing temporal expressions

- Map temporal expressions to specific time points or intervals.
- Encode time information according to a standard (ISO 8601)

<table>
<thead>
<tr>
<th>Unit</th>
<th>Pattern</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully specified dates</td>
<td>YYYY-MM-DD</td>
<td>2012-04-30</td>
</tr>
<tr>
<td>Weeks</td>
<td>YYYY-nnW</td>
<td>2012-19W</td>
</tr>
<tr>
<td>Clock times</td>
<td>HH:MM:SS</td>
<td>03:14:15</td>
</tr>
<tr>
<td>Dates and times</td>
<td>YYYY-MM-DDTHH:MM:SS</td>
<td>2012-04-30T03:14:15</td>
</tr>
<tr>
<td>Financial quarters</td>
<td>YYYY-Qn</td>
<td>2012-Q2</td>
</tr>
</tbody>
</table>
Normalizing temporal expressions

anchor selection connect each relative TE with an absolute TE

- recomputes relative time to the document creation date (CR_DATE)
- connects relative time to the nearest time expression with compatible granularity (PR_DATE)

date normalization

- absolute TES – translate to representation standard
- relative TEs – use the anchor, relative position to the anchor, distance from anchor

\[ \text{two years later} \rightarrow \text{ANCHOR} + 2 \]

He started studying on March 30 2004, and passed the exam the following Friday.
Issues in normalization

- embedded time expressions: "the eve of the new year, sixty years ago today"
- reported speech: "He concluded the 1998 annual meeting saying: 'The next year will be the eve of a new era for our company'"
Events

STATIVES  *know, sit, be clever, be happy* ...
ACTIVITIES  *walk, run, talk, march, paint* ...
ACCOMPLISHMENTS  *build, cook, destroy* ...
ACHIEVEMENTS  *notice, win, blink, find, reach* ...

Events have an implicit temporal dimension
Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip... American Airlines, immediately matched the move, spokesman Tim Wagner said. ...
Event analysis

• determine event structure (event subclasses and parts, participants)
• analyse temporal dimension:
  • tense – indicates location of event in time, via verb inflections, modals, auxiliaries, etc.
  • grammatical aspect – indicates whether event is ongoing, finished, completed
  • time adverbials – indicate relations between events and/or times and temporal relations
Event detection and analysis

Identify mentions of events in text:

- verbs: cite, say, increase, ...
  but not all: have, take, have, ... (in certain contexts)
- nouns: move, increase, ...

Commonly used features:

- affixes
- nominalization suffix
- part of speech
- light verb
- subject syntactic category
- morphological stem
- verb root
- WordNet hypernyms

prefixes and suffixes of the target word
e.g. -tion
part of speech of the target word
whether the target is governed by a light verb
noun, pronoun, noun phrase, ...
stemmed version of the target word
root form of the verb basis if the target is a nominalized verb
Hypernym set for the target
Temporal event analysis

- connect events to temporal expressions
- establish relative positions of events on the time axis
- map events onto a timeline
Event temporal relations

- A is before B or B is after A
- A meets B or B is met by A
- A overlaps with B or B is overlapped by A
- A starts B or B is started-by A
- A during B or B contains A
- A finishes B or B is finished-by A
- A and B are cotemporal
Mary left on Thursday and John arrived the day after.

Mary left on

<TIMEX3 tid="t1" type="DATE" value="1998-WXX-4" temporalFunction="true" anchorTimeID="t0"> Thursday </TIMEX3>

and John arrived

<TIMEX3 tid="t2" type="DATE" value="1998-WXX-5" temporalFunction="true" anchorTimeID="t1"> the day </TIMEX3>
after.
Task for next week

1. Read the TimeBank annotation guidelines
2. Have a look at the posted data