Advanced Topics in Information Retrieval

Temporal Information Extraction

Vinay Setty
vsetty@mpi-inf.mpg.de

Jannik Strötgen
jannik.stroetgen@mpi-inf.mpg.de

ATIR – June 16, 2016
Why is **temporal information** crucial for information retrieval?
Time in queries

temporal information needs are frequent

query log analyses

- 1.5% queries with **explicit temporal intent** [Nunes et al. 2008]
- 7% queries with **implicit temporal intent** [Metzler et al. 2009]
- 13.8% **explicit**, 17.1% **implicit** [Zhang et al. 2010]

different types of temporal information in IR

- time as **dimension of relevance**
- time as **query topic**
Gedankenexperiment

What did Alexander von Humboldt do between *late 18th century* and *early 19th century* in Latein America?
Let’s search ...

alexander von humboldt latin america late 18th century to early 19th century

Snippets tell us a lot...

Alexander von Humboldt - Wikipedia, the free encyclopedia

Jump to Achievements of the Latin American expedition - Friedrich Wilhelm Heinrich Alexander von Humboldt was ..... general public until the late nineteenth century, in the case of the .... dealt with the early voyages to the Americas, pursuing his interest ... family groupings in the eighteenth century, showing father of ...

Humboldt's Cosmos: Alexander von Humboldt and the Latin
www.amazon.com › ... › Historical

Amazon.com: Humboldt’s Cosmos: Alexander von Humboldt and the Latin American Journey ... Humboldt was in his late twenties, a German aristocrat of independent means, .... systems, this book will reawaken your appreciation of this 19th century. ... An amazing journey of discovery in South America, Cuba, and Mexico.
Let’s search ...

terms occurring in the query

Alexander von Humboldt
Latin American
Wilhelm Heinrich
early
late nineteenth century
eighteenth century
Humboldt's Cosmos: Alexander von Humboldt and the Latin
American Historical Campaign's Amazon.com
Amazon.com
Journey ... Humboldt was in his late entries, a German aristocrat of independent means,
Let’s search ...

alexander von humboldt latin america late 18th century to early 19th century

not highlighted:
expressions matching query interval / region

Alexander von Humboldt - Wikipedia, the free encyclopedia
Jump to Achievements of the Latin American expedition - Friedrich Wilhelm Heinrich Alexander von Humboldt was ..... general public until the late nineteenth century, in the case of the .... dealt with the early voyages to the Americas, pursuing his interest ... family groupings in the eighteenth century, showing father of ...

Humboldt's Cosmos: Alexander von Humboldt and the Latin American Journey ... Humboldt was in his late twenties, a German aristocrat of independent means, .... systems, this book will reawaken ... An amazing journey of discovery in South America, Cuba, Mexico.
Excerpt of the Wikipedia page *Alexander von Humboldt*.
Problems of standard IR approaches

**temporal and geographic expressions**
- (seem to be) treated as regular terms
- semantics is lost

→ **should be extracted and normalized**

**query functionality**
- how to search for time intervals?
- how to search for geographic regions?

→ **should be defined and provided**

**results**
- same ranking as for standard text search
- no time-/geo-centric exploration features

→ **special ranking is required**

→ **time-/geo-centric exploration should be possible**
Things that need to be done

<table>
<thead>
<tr>
<th>next week</th>
</tr>
</thead>
<tbody>
<tr>
<td>temporal information retrieval</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>today</th>
</tr>
</thead>
<tbody>
<tr>
<td>temporal information extraction</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>maybe later</th>
</tr>
</thead>
<tbody>
<tr>
<td>geographic and event-centric info</td>
</tr>
</tbody>
</table>

© Jannik Strötgen – ATIR-07
Outline

1. Temporal Information
2. Temporal Tagging
3. Evaluation
4. HeidelTime
5. Temponym tagging
6. NLP Pipeline Architectures
Outline

1. Temporal Information
2. Temporal Tagging
3. Evaluation
4. HeidelTime
5. Temponym tagging
6. NLP Pipeline Architectures
Is time that important?

Temporal information plays an important role in many types of text documents.
Is time that important?

Temporal information has important key characteristics:
- expressions can be compared

Examples:
- before: 2010 / 2016
- overlap: 1960s / 1955 to 1965
- during: June 2016 / 2016
- ...

Allen’s interval algebra
[Allen 1983]
Given two intervals X and Y, one of 13 relations holds between them
Is time that important?

Temporal information has important key characteristics:

Temporal information is **well-defined**:

- expressions can be **compared** with each other

1) X before Y
2) X equal Y
3) X meets Y
4) X overlaps Y
5) X during Y
6) X starts Y
7) X finishes Y

Source: [Strötgen & Gertz 2016]
Is time that important?

Temporal information has important key characteristics:

Temporal information can be **normalized**:
- expressions with **same semantics** → **same value**

Examples:
- June 16, 2016
- today
- heute, aujourd’hui, hoy, oggi, ...

→ 2016-06-16

→ Temporal information is **term- and language-independent**
Is time that important?

Temporal information has important key characteristics.

Temporal information can be normalized:
- expressions with **same semantics** → **same value**

Source: [Strötgen & Gertz 2016]
Is time that important?

Temporal information has important key characteristics:
- expressions of different granularities

Temporal information can be organized hierarchically:

```
  ...  2014  2015  2016  ...
    /    /  /   /
   /     /  /    /
  /       /  /     /
 /         /  /      \
/          /  /       \
/          /  /        \
/          /  /         \
/          /  /          \
/          /  /           \
|          |  /             \
|          |  /              \
|          |  /               \
|          |  /                \
2015-03-11 2015-03-12 2015-03 ...
```

© Jannik Strötgen – ATIR-07
Is time that important?

Source: [Strötgen & Gertz 2016]
Temporal Tagging

temporal expressions

- a special type of “named entity”
- extraction sometimes covered by NER tools
- intuitively: normalization is very important
Outline

1. Temporal Information
2. Temporal Tagging
3. Evaluation
4. HeidelTime
5. Temponym tagging
6. NLP Pipeline Architectures
Temporal Tagging

the two tasks of temporal taggers

1. *extraction* of temporal expressions

Greece Makes ‘Good Progress’ in Payment Talks

By Maria Petrouli and Natalie Petrouli

Greek Finance Minister Evangelos Venizelos made “good progress” in a second round of talks with the European Union and International Monetary Fund aimed at staying off default, the EU said.

The telephone meeting followed a meeting that were intended to calm concerns that Greece's reduction targets and to clear the way for a sixth installment of rescue funds. The EU statement said a “full mission” will return to Athens for discussions in coming days at the IMF’s annual meeting in Washington.

Staying in the euro area is an “irreversible and fundamental national choice,” Venizelos said in a statement. “We’ve demonstrated that we have a solid data and economic foundation to be a part of the eurozone, determined to tackle once and for all.”

The EL comes as Greece payment for Greece is on track to be released. Finance Minister George Papandreou fight...
Temporal Tagging

the two tasks of temporal taggers

1. **extraction** of temporal expressions
2. **normalization** of temporal expressions

**Greece Makes ‘Good Progress’ in Payment Talks**

By Maria Petrakis and Natalie

Sep 20, 2011 10:39 PM GMT+0200

Greek Finance Minister Evangelos Venizelos made “good progress” in a second round of talks with the European Union and International Monetary Fund aimed at staying off default, the EU said.

The telephone meeting, which followed an earlier discussion in coming days at the IMF’s annual meeting in Washington.

Staying in the euro area is an “irreversible and fundamental national choice,” Venizelos said in a statement today. He said Greece is “determined to tackle once and for all.”

The EC’s statement follows a meeting of eurozone finance ministers in a day of meetings to discuss Greece’s fate in the currency union.

The EC statement said the country is “on track to be released from its third bailout.”

**Normalization of relative and underspecified expressions**

- **tonight** → 2011-09-20TNI
- **yesterday** → 2011-09-19
- **next week** → 2011-W39
- **Sept. 20, 2011** → 2011-09-20
- **next month** → 2011-10
Temporal Expressions

different types of temporal expressions

temporal markup language TimeML defines four types:
[Pustejovsky et al. 2005] (http://timeml.org/)

- Dates
  → June 24, 2013
  → September 2000
  → two weeks ago

- Times
  → 3 p.m.
  → yesterday morning
  → 2012-06-28T16:25

- Durations
  → two weeks
  → 12.5 hours
  → several months

- Sets
  → every day
  → annually
  → twice a month

dates and times particularly valuable for IR
Temporal Expressions

- **explicit**
  - June 24, 2013
  - the 20th century
  - easy to normalize

- **implicit**
  - Christmas 2012
  - Columbus Day 2006
  - additional knowledge

- **relative**
  - two weeks ago
  - yesterday
  - reference time

- **underspecified**
  - Monday
  - June 24
  - reference time and relation to it

**main challenge for temporal taggers**

normalization of relative and underspecified expressions
Normalization of temporal expressions

main challenge

normalization of relative and underspecified expressions

Document Creation Time: 2000-12-26

...On **Thursday**, the Census Bureau will publish the official population count for the United States, including the state-by-state totals required under the Constitution to determine how many seats each state is allocated in the House. The figures, eagerly awaited by many state government officials, are the first in a wave of releases of demographic data based on the 2000 census. ... Population estimates issued periodically by the Census Bureau indicate that as of **October**, 275,843,000 people were living in ... Additional seats are then assigned to each state based on a person-to-House-member ratio that changes **every 10 years** because the country’s population keeps growing ...
Normalization of temporal expressions

**main challenge**

normalization of relative and underspecified expressions

---

Document Creation Time: **2000-12-26**

… On **Thursday**, the Census Bureau will publish the official population count for the United States, including the state-by-state totals required under the Constitution to determine how many seats each state is allocated in the House. The figures, eagerly awaited by many state government officials, are the first in a wave of releases of demographic data based on the **2000** census. …

Population estimates issued periodically by the Census Bureau indicate that as of **October**, 275,843,000 people **were living** in … Additional seats are then assigned to each state based on a person-to-House-member ratio that changes **every 10 years** because the country’s population keeps growing …
Normalization of temporal expressions

temporal tagging of news articles

document creation time is important

Document Creation Time: 2000-12-26

... On Thursday, the Census Bureau will publish the official population count for the United States, including the state-by-state totals required under the Constitution to determine how many seats each state is allocated in the House. The figures, eagerly awaited by many state government officials, are the first in a wave of releases of demographic data based on the 2000 census. ... Population estimates issued periodically by the Census Bureau indicate that as of October, 275,843,000 people were living in ... Additional seats are then assigned to each state based on a person-to-House-member ratio that changes every 10 years because the country’s population keeps growing ...
Temporal expressions in various corpora

TimeBank corpus [Pustejovsky et al. 2003]
news articles with manually annotated temporal expressions

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Duration</th>
<th>Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occurrences [%]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

© Jannik Strötgen – ATIR-07
Temporal expressions in various corpora

TimeBank corpus [Pustejovsky et al. 2003]
news articles with manually annotated temporal expressions

Occurrences [%] (dates, times)

© Jannik Strötgen – ATIR-07
Temporal Tagging of News Articles

characteristics
- document creation time (DCT) plays a crucial role
- many date expressions
- many relative and underspecified expressions

challenges
- detection of relations between reference time and underspecified expressions
- detection of reference times for relative expressions where DCT is not the reference time

examples
- news articles
- letters, formal emails, etc.
Temporal Tagging of News Articles

**most research** on temporal tagging focused on processing of (English) **news articles**

- manually annotated corpora, e.g.,
  - TimeBank [Pustejovsky et al. 2003]
- research competitions, e.g.,
  - TempEval series e.g., [UzZaman et al. 2013]
- temporal taggers, e.g.,
  - GUTime [Verhagen et al. 2005]

**different domains** pose different challenges
Normalization of temporal expressions

narrative documents
reference time has to be detected in the text

1979: Soviet invasion

On December 7, 1979, Soviet informants to the Afghan Armed Forces ... and began to land in Kabul on December 25. On December 27, 1979, 700 Soviet troops dressed in Afghan uniforms, ... That operation began at 19:00 hr., ... The operation was fully complete by the morning of December 28, 1979. ... According to the Soviet Politburo they were complying with the 1978 Treaty of Friendship, .......... Soviet ground forces, under the command of Marshal Sergei Sokolov, entered Afghanistan from the north on December 27. In the morning, the 103rd Guards 'Vitebsk' ...
Temporal expressions in various corpora

WikiWars corpus [Mazur & Dale 2010]
Wikipedia articles with manually annotated temporal expressions
Temporal expressions in various corpora

WikiWars corpus [Mazur & Dale 2010]
Wikipedia articles with manually annotated temporal expressions
Temporal Tagging of Narrative Documents

characteristics

- independent of document creation time
- many explicit expressions
- often long texts with complex temporal discourse structure

challenges

- reference time detection for relative and underspecified expressions
- normalization of expressions referring to historic dates

examples

- Wikipedia articles
- descriptive documents, biographies, documents about history, etc.
Temporal Tagging of Colloquial Texts

- relation to reference time
- non-standard language (errors, word creations, ...)
- missing context information

---

**SMS** 2010-01-10T05:19
What's it u wanted 2 say last nite?

**SMS** 2010-09-23T09:50
Yo! Rem to come for lab tmr :-)

**SMS** 2011-02-16T12:42
... andy is available at **10 am** in his office
Temporal expressions in various corpora

Time4SMS corpus [Strötgen & Gertz 2012]
short messages with manually annotated temporal expressions
Temporal expressions in various corpora

**Time4SMS corpus** [Strötgen & Gertz 2012]

short messages with manually annotated temporal expressions

---

**Occurrences [%]**

- **(dates, times)**
  - news
  - narrative
  - colloquial

---

**Occurrences [%]**

- **(dates, times)**
  - explicit
  - implicit
  - relative, ref=act
  - relative, ref≠act
  - underspec., ref=act
  - underspec., ref≠act
  - unresolv.
Temporal Tagging of Colloquial Texts

characteristics

- use of “noisy” language
- rarely any explicit expressions
- document creation time plays a crucial role

challenges

- spelling variations and non-standard vocabulary
- detection of relation between reference time and underspecified expressions
- missing context information

examples

- short messages, tweets, social media content, etc.
Temporal Tagging of Autonomic Texts

... Subjects consumed one tablet per day containing ...
... Subjects were assessed at baseline, three and six months ...
... Clinical pathology analysis was performed at baseline and six months later ...

- often no real reference time
- local semantics (document time frame)
- “time point zero”
Temporal expressions in various corpora

**Time4SCI corpus** [Strötgen & Gertz 2012]
clinical trials with manually annotated temporal expressions

---

**Occurrences [%]**

- **Date**: news (red), narrative (red crossed), colloquial (blue), scientific (blue crossed)
- **Time**: news (red), narrative (red crossed), colloquial (blue), scientific (blue crossed)
- **Duration**: news (red), narrative (red crossed), colloquial (blue), scientific (blue crossed)
- **Set**: news (red), narrative (red crossed), colloquial (blue), scientific (blue crossed)
Temporal expressions in various corpora

**Time4SCI corpus [Strötgen & Gertz 2012]**
clinical trials with manually annotated temporal expressions
Temporal Tagging of Autonomic Texts

characteristics
- local (autonomic) time frame
- unresolvable relative and underspecified expressions

challenges
- validity of local time frame
- time point zero detection

examples
- clinical trials, clinical descriptions
- literary texts, etc.
Approaches

extraction task
- rule-based
- machine learning
- semantic parsing
- hybrid

normalization task
- rule-based
- hybrid
Machine learning-based extraction

a typical classification problem:

- **IOB classification**
- input: sequence of tokens
- decide for each token if it is inside (I), outside (O) or the beginning (B) of a temporal expressions

In March, I finished my PhD which I started two years ago.

O B O O O O O O O O O B I I O
Machine learning-based extraction

frequently used classifier
- maximum entropy
- support vector machines
- conditional random fields

typically used features
- **lexical features** (part-of-speech, token, character-based, lists)
- **syntactic features** (base phrase chunks)
- semantic features (semantic role labels)
- external features (information of other temporal taggers)

learning based on training data
Rule-based extraction

features and techniques

- pattern files
- regular expressions
- part-of-speech information
- positive and negative rules
- cascaded organization of rules
Rule-based extraction

temporal tagging vs. standard NER

- divergence of temporal expressions is very limited
- the number of persons and organizations and variety of names referring to these entities probably infinite

rules for extraction can be used for normalization
Normalization

**rule based approaches**

- normalization information for patterns
- reference time detection (DCT, previous expression)
- relation to reference time → domain-dependent
- news domain: tense information can be helpful
- narrative domain: chronology assumption (for short passages between underspecified expressions and reference times)
Temporal taggers

- **SUTime** [Chang & Manning 2012, 2013]
- **HeidelTime** [Strötgen & Gertz 2010, 2013; Strötgen et al. 2013]
- **ClearTK-TimeML with Timenorm** [Bethard 2013]
Outline

1. Temporal Information
2. Temporal Tagging
3. Evaluation
4. HeidelTime
5. Temponym tagging
6. NLP Pipeline Architectures
Evaluation

**extraction task**

- TP: annotated by the system and in the gold standard
- FP: annotated by the system but not in the gold standard
- TN: neither annotated by the system nor in the gold standard
- FN: not annotated by the system but in the gold standard

<table>
<thead>
<tr>
<th>system prediction</th>
<th>gold standard (ground truth)</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive</td>
<td>TP</td>
</tr>
<tr>
<td>negative</td>
<td>FN</td>
</tr>
<tr>
<td>positive</td>
<td>FP</td>
</tr>
<tr>
<td>negative</td>
<td>TN</td>
</tr>
</tbody>
</table>
Evaluation

<table>
<thead>
<tr>
<th>system prediction</th>
<th>gold standard (ground truth)</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive</td>
<td>positive</td>
</tr>
<tr>
<td>negative</td>
<td>negative</td>
</tr>
<tr>
<td></td>
<td>TP</td>
</tr>
<tr>
<td></td>
<td>FN</td>
</tr>
<tr>
<td></td>
<td>FP</td>
</tr>
<tr>
<td></td>
<td>TN</td>
</tr>
</tbody>
</table>

measures

\[ p = \frac{TP}{TP + FP} \]
\[ r = \frac{TP}{TP + FN} \]
\[ f_1 = \frac{2 \cdot p \cdot r}{p + r} \]
Evaluation

example

In March, I finished . . . I started about two years ago.

gold: <TIMEX3>March</TIMEX3>
< TIMEX3 >about two years ago</TIMEX3>

system: <TIMEX3>March</TIMEX3>
< TIMEX3 >two years ago</TIMEX3>

strict matching

TP = 1
FP = 1
FN = 1

relaxed matching

TP = 2
FP = 0
FN = 0
Evaluation

normalization task

**normalization accuracy**
how many of the correctly extracted expressions are also normalized correctly?

- not directly comparable between systems
- depends on recall in extraction task

**value f1 score**
TP: correctly extracted and correctly normalized

- combined score for extraction and normalization
- most widely used

© Jannik Strötgen – ATIR-07
Evaluation

example

In March, I finished ... I started about two years ago.

gold: <TIMEX3 value="2016-03">March</TIMEX3> <TIMEX3 value="2014-06-16">about two years ago</TIMEX3>
system: <TIMEX3 value="2016-03">March</TIMEX3> <TIMEX3 value="2014-06-16">two years ago</TIMEX3>

value $f_1$ strict matching
- TP = 1
- FP = 1
- FN = 1

value $f_1$ relaxed matching
- TP = 2
- FP = 0
- FN = 0

most meaningful value $f_1$ relaxed matching
Evaluation Campaigns

TempEval competitions: [Verhagen et al. 2010; UzZaman et al. 2013]

**goal of TempEval**

temporal information extraction and push the field forward!

**procedure**
- provide training data (manually annotated corpora)
- promote the task
- make researchers participate, let them develop a system
- evaluate systems with test data (held-out gold standard)
- compare the systems’ performance, see what worked

**subtasks in TempEval all based on TimeML**
- temporal tagging
- event extraction
- temporal relation extraction
Evaluation Campaigns

Temporal tagging at TempEval

- news corpora only
- organizers concluded: “that rule-engineering and machine learning are equally good at timex recognition”

2015 / 2016: Clinical TempEval

- clinical texts
- temporal tagging subtask with extraction only
Outline

1. Temporal Information
2. Temporal Tagging
3. Evaluation
4. HeidelTime
5. Temponym tagging
6. NLP Pipeline Architectures
HeidelTime

HeidelTime [Strötgen & Gertz 2010, 2013]

- rule-based, multilingual, cross-domain temporal tagger

Extraction

- mainly based on regular expressions
- linguistic features (POS, POS of next token, ...)
- knowledge resources (names of months, holidays, ...)

Normalization

- linguistic clues (tense in sentence, ...)
- domain-specific normalization strategies
HeidelTime’s Architecture

- domain-specific normalization strategies
- resource interpreter
  → language-independent

- patterns
- normalization knowledge
- rules
  → language-dependent

HeidelTime has a well-defined rule syntax
HeidelTime’s Language Resources

Pattern files:
- frequently used terms

Normalization files:
- contain normalized values of such terms
HeidelTime’s Language Resources

After read by resource interpreter, accessible by rules:

\[
\text{reMonthLong} = (\text{January}|\text{February}|...)
\]
\[
\text{reMonthShort} = (\text{Jan}.?|\text{Feb}.?|\text{Mar}.?|...)
\]
\[
\text{reMonthNumber} = (10|11|12|0?[1-9])
\]
\[
\text{normMonth}("\text{January}" ) = "01"
\]
\[
\text{normMonth}("\text{Jan}."") = "01"
\]
\[
\text{normMonth}("\text{Jan}" ) = "01"
\]
\[
\text{normMonth}("01") = "01"
\]
\[
\text{normMonth}("1") = "01"
\]

Rule files:
- every rule contains at least:
  (i) rule name, (ii) extraction part, (iii) value normalization part

Details below...
HeidelTime – Simple Rule Example

A rule for **January 8, 2010**:  

RULE_NAME="date_r1"
EXTRACTION="%reMonthLong %reDayNumber, %reYear4Digit"
NORM_VALUE="group(3)-%normMonth(group(1))-%normDay(group(2))"
HeidelTime – Simple Rule Example

A rule for **January 8, 2010:**

RULE_NAME=“date_r1”
EXTRACTION=“%reMonthLong %reDayNumber, %reYear4Digit”

January 8, 2010

NORM_VALUE=“group(3)-%normMonth(group(1))-%normDay(group(2))”

=“2010-%normMonth(January)-%normDay(8)”

=“2010-01-08”

What about more difficult expressions?
HeidelTime – More Complex Rule Example

How to normalize underspecified and relative expressions?

A rule for **November 21st**:

RULE_NAME="date_r2"
EXTRACTION="(%reMonthLong|%reMonthShort) " +
        "(%reDayNumberTh|%reDayNumber)"
NORM_VALUE="UNDEF-year-%normMonth(group(1))-%normDay(group(4))"

**Example:**
- Extracted expression: “**November 21st**”
- NORM_VALUE=“**UNDEF-year-11-21**”
HeidelTime – More Complex Rule Example

Example:
- Extracted expression: “November 21st”
- NORM_VALUE=“UNDEF-year-11-21”

Normalization:
- rules use “UNDEF”-expressions
- disambiguation in the source code (domain-dependent)

Normalization of “UNDEF-year-11-21” (simplified)
- News: document creation time and tense in sentence
- Narrative: previously mentioned expression, chronology assumption
HeidelTime – More Complex Rule Example

- **more constraints can be added**
  → e.g., part of speech constraints

- **negative rules can be added**
  → to prevent wrong expressions from being tagged as temporal expressions

- “In **2000**, a new era begins”
- “In **2000** miles, a new area begins”

**Several further things to specify, e.g.,**
- further normalization information
- “random” tokens
HeidelTime

4 domains
- news, narrative, colloquial, autonomic

13 languages
- en, de, es, it
- 4 developed by colleagues (ar, vn, cn, est)
- 5 developed at other institutes (fr, ru, du, hr, pt)


easy-to-extend to further languages

publicly available
- widely used in the research community
- first domain-sensitive temporal tagger
- only tagger for some of the languages
HeidelTime – Extensive Evaluation

the value of HeidelTime’s domain-sensitive strategies

cross-domain evaluation [Strötgen & Gertz 2012]

<table>
<thead>
<tr>
<th>corpus</th>
<th>strategy</th>
<th>extraction</th>
<th>normalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>news</td>
<td>news</td>
<td>91.1</td>
<td>78.6</td>
</tr>
<tr>
<td></td>
<td>narratives</td>
<td>91.1</td>
<td>61.5</td>
</tr>
<tr>
<td>narrative</td>
<td>news</td>
<td>87.9</td>
<td>56.9</td>
</tr>
<tr>
<td></td>
<td>narratives</td>
<td>87.9</td>
<td>78.7</td>
</tr>
</tbody>
</table>

Don’t trust a tagger developed for news if you want to process narratives (e.g., Wikipedia)
HeidelTime is Publicly Available

Publicly available:
- UIMA version
- as Gate plugin (GATE-Time)
- standalone version (Java)
- online demo

feedback is appreciated!
you’ll (have to) use it in your assignment
Developing Language Resources Manually

**Spanish resource development** (in the context of TempEval-3)

- Translation of pattern files
- Translation of normalization files
- Iterative rule development
  1. starting with (simple) English rules
  2. checking Spanish training data for errors: partial matches, false positives, false negatives, incorrect normalizations
  3. adapting pattern and normalization files where necessary; adapting/adding rules to improve results on training data
     → until results could not be improved anymore
Goal: Temporal Tagging of All Languages

so far:
- manual resource development for each added language

disadvantages:
- labor intensive
- time intensive
- language knowledge required
- there are many more languages not yet addressed

now: first step towards temporal tagging of all languages
Developing Language Resources Automatically

HeidelTime 2.0 approach [Strötgen & Gertz 2015]

- language-independent resources
  - some patterns and normalization information are valid for all (many) languages, e.g., numbers for days and months
- simplified English resources as starting point for translations
  - only normalization files
  - without regular expressions
  - for each context separately
  - e.g., normMonthLong containing:
    - “January”, “01”
    - “February”, “02”
    - ...

- resource development process for “all languages”
- language-independent rules
Developing Language Resources Automatically

Resource development process for “all languages”

Simplified English resources

```
normMonthLong
"January","01"
"February","02"
...
```

extract patterns

```
January
February
...
```

for each language

\[
\text{January translations = \{ 
  \text{german: Januar}, 
  \text{spanish: enero} 
\}}
\]

add patterns

\[
\text{February translations = \{ 
  \text{german: Februar}, 
  \text{spanish: febrero} 
\}}
\]

add normalizations

\[
\text{normMonthLong}
"January","01"
"February","02"
\]

\[
\text{reMonthLong}
"January","01"
"February","02"
\]

\[
\text{german}
Januar
Februar
\]

\[
\text{spanish}
enero
febrero
\]

\[
\text{wiktionary}
\]

© Jannik Strötgen – ATIR-07
Developing Language Resources Automatically

Language-independent rules

- rules without any language-dependent tokens (words)
- based on original English rules only
- add “creative rules”
- allow for fuzzy matching of patterns (to avoid problems with morphology-rich languages)

Assumption

- obviously, not all rules required for all languages
  but: “unnecessary” rules are unlikely to harm results
## Evaluation

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ar: Arabic test-50*</td>
<td>90.9</td>
<td>90.9</td>
<td>90.9</td>
<td>82.2</td>
<td>90.4</td>
<td>91.7</td>
<td>31.8</td>
<td>47.2</td>
<td>38.0</td>
<td>80.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ch: TE-2 test impro.</td>
<td>95.8</td>
<td>89.3</td>
<td>92.4</td>
<td>79.5</td>
<td>86.0</td>
<td>100</td>
<td>9.5</td>
<td>17.3</td>
<td>7.6</td>
<td>44.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hr: WikiWarsHR</td>
<td>92.6</td>
<td>90.5</td>
<td>91.5</td>
<td>80.8</td>
<td>88.3</td>
<td>87.3</td>
<td>6.8</td>
<td>12.6</td>
<td>9.7</td>
<td>77.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fr: FR-TimeBank</td>
<td>91.9</td>
<td>90.1</td>
<td>91.0</td>
<td>73.6</td>
<td>80.9</td>
<td>87.2</td>
<td>59.5</td>
<td>70.8</td>
<td>54.6</td>
<td>77.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>de: WikiWarsDE</td>
<td>98.7</td>
<td>89.3</td>
<td>93.8</td>
<td>83.0</td>
<td>88.5</td>
<td>98.4</td>
<td>64.7</td>
<td>78.1</td>
<td>59.7</td>
<td>76.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>it: EVALITA’14 test</td>
<td>92.7</td>
<td>86.1</td>
<td>89.3</td>
<td>75.0</td>
<td>84.0</td>
<td>98.5</td>
<td>41.2</td>
<td>58.1</td>
<td>49.3</td>
<td>84.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>es: TempEval-3 test</td>
<td>96.0</td>
<td>84.9</td>
<td>90.1</td>
<td>85.3</td>
<td>94.7</td>
<td>95.5</td>
<td>53.8</td>
<td>68.8</td>
<td>58.5</td>
<td>85.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vn: WikiWarsVN</td>
<td>98.2</td>
<td>98.2</td>
<td>98.2</td>
<td>91.4</td>
<td>93.1</td>
<td>84.0</td>
<td>45.5</td>
<td>59.0</td>
<td>27.1</td>
<td>45.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pt: PT-TimeBank test</td>
<td>87.3</td>
<td>75.9</td>
<td>81.2</td>
<td>63.5</td>
<td>78.2</td>
<td>91.5</td>
<td>59.3</td>
<td>72.0</td>
<td>59.4</td>
<td>82.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pt: PT-TimeBank train</td>
<td>83.3</td>
<td>73.1</td>
<td>77.9</td>
<td>54.5</td>
<td>70.0</td>
<td>88.2</td>
<td>51.0</td>
<td>64.6</td>
<td>50.4</td>
<td>78.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ro: Ro-TimeBank</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>31.9</td>
<td>11.4</td>
<td>16.9</td>
<td>7.8</td>
<td>46.2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Evaluation

On publicly available corpora
- compare automatically generated resources with HeidelTime’s manually created resources
- worse, but very promising results (for many languages)

Before
- HeidelTime supported 13 languages
  and no other temporal taggers for other languages available

HeidelTime 2.0:
- HeidelTime as baseline tagger for 200+ languages
- automatically created resources as a starting point
Outline

1. Temporal Information
2. Temporal Tagging
3. Evaluation
4. HeidelTime
5. Temponym tagging
6. NLP Pipeline Architectures
Outline

1. Temporal Information
2. Temporal Tagging
3. Evaluation
4. HeidelTime
5. Temponym tagging
6. NLP Pipeline Architectures
NLP Pipeline Architectures

NLP tasks can often be split into multiple sub-tasks

- e.g., dependency parsing:
  - sentence splitting
  - tokenization
  - part-of-speech tagging
  - parsing

- pre-processing of corpora, e.g., for semantic search
- UIMA https://uima.apache.org/
- GATE https://gate.ac.uk/
- NLTK http://www.nltk.org/
- Stanford CoreNLP http://stanfordnlp.github.io/CoreNLP/
The Pipeline Principle – Why a (UIMA) Pipeline

UIMA: Unstructured Information Management Architecture

- component framework for unstructured data
- helps to combine tools not built to be used together
- data structure: Common Analysis Structure (CAS)
The Pipeline Principle – Why a (UIMA) Pipeline

3 Types of Components
- collection readers
- analysis engines
- CAS consumer
The Pipeline Principle – Why a (UIMA) Pipeline

Collection Reader
- reads documents from a source (e.g., file system, database)
- creates a CAS object for each document
- adds first annotations, e.g., document text, metadata
The Pipeline Principle – Why a (UIMA) Pipeline

Analysis Engines

- usually several analysis engines
The Pipeline Principle – Why a (UIMA) Pipeline

**Analysis Engines**
- read the CAS
- analyze the documents (document text)
- add annotations to the CAS
Cas Consumer

- reads the CAS
- perform final processing (indexing, evaluation, ...)
- output annotations
The Pipeline Principle – Why a (UIMA) Pipeline

What’s the clue?

- single components are not directly connected
- “connected” via CAS
Summary

- **Time** is important and has many nice characteristics (it can be normalized!)
- **Temporal Tagging** extraction and normalization of temporal expressions
- Differences between various types of documents: domain-sensitive temporal tagging is crucial
- Several approaches to temporal tagging
- **HeidelTime**: multilingual and domain-sensitive
- **Temponyms**: postponed to next week

Thank you for your attention!
More Information on Temporal Tagging

Book on temporal tagging:
References

mentioned in the slides:

Nunes et al. 2008: Use of Temporal Expressions in Web Search, ECIR.
Metzler et al. 2009: Improving Search Relevance for Implicitly Temporal Queries, SIGIR.
Zhang et al. 2010: Learning Recurrent Event Queries for Web Search, EMNLP.
Allen 1983: Maintaining Knowledge about Temporal Intervals, Comm. of the ACM.
Verhagen et al. 2005: Automating Temporal Annotation with TARSQI, ACL.
Mazur & Dale 2010: WikiWars: A New Corpus for Research on Temporal Expressions, EMNLP.
Strötgen & Gertz 2012: Temporal Tagging on Different Domains: Challenges, Strategies, and Gold Standards, LREC.
Chang & Manning 2012: SUTime: A Library for Recognizing and Normalizing Time Expressions, LREC.
Strötgen & Gertz 2013: Multilingual and Cross-domain Temporal Tagging, LRE journal.
Moriceau & Tannier 2014: French Resources for Extraction and Normalization of Temporal Expressions with HeidelTime, LREC.
Camp & Christiansen 2012: Resolving Relative Time Expressions in Dutch Text with Constraint Handling Rules, CSLP.
Strötgen & Gertz 2015: A Baseline Temporal Tagger for All Languages, EMNLP.
Kuzey et al. 2016a: As Time Goes By: Comprehensive Tagging of Textual Phrases with Temporal Scopes, WWW.