



max planck institut
informatik

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

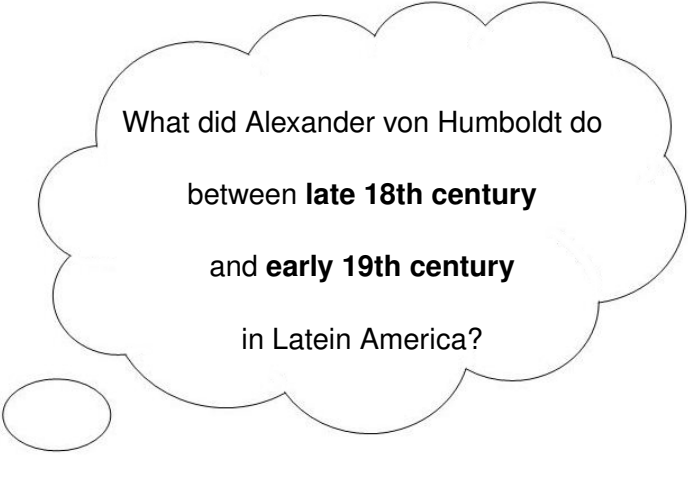
- 1.5% queries with **explicit temporal intent** [Nunes et al. 2008]
- 7% queries with **implicit temporal intent** [Metzler et al. 2009]
- 13.8% **explicit**, 17.1% **implicit** [Zhang et al. 2010]

different types of temporal information in IR

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

more next week

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...

About 91,000 results (1.16 seconds)

Alexander von Humboldt - Wikipedia, the free encyclopedia

https://en.wikipedia.org/wiki/Alexander_von_Humboldt ▼ Wikipedia ▼

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, Inc. ▼

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 ...

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



highlighted:

terms occurring in the query

About 91,000 results (1.16 seconds)

Alexander von Humboldt
[https://en.wikipedia.org/wiki/Alexander_von Humboldt](https://en.wikipedia.org/wiki/Alexander_von_Humboldt)
 ... Jump to Achievements of the Alexander von Humboldt
 the case of the dealt with the family groupings in the
 public until the s to the Amer
 Wilhelm Heinrich
 early late nineteenth century
 eighteenth century
 Humboldt's Cosmos: Alexander von Humboldt and the Latin
www.amazon.com
 Amazon.cc Alexander von Humboldt and the Latin American
 Journey ... Humboldt was in his late enties, a German aristocrat of independent means,
 systems, this book
 journal of discovery in latin america
 19th century
 © Jannik Strötgen – ATIR-07



Let's search ...

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



not highlighted:

expressions matching query interval / region

About 91,000 results (1.16 seconds)

Alexander von Humboldt - Wikipedia, the free encyclopedia

https://en.wikipedia.org/wiki/Alexander_von_Humboldt ▼ Wikipedia ▼

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, Inc. ▼

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

journey of discovery in South America
informatik

Cuba, Mexico

© Jannik Strötgen – ATIR-07



Improved snippets

expressions matching query interval / region



portrait of Humboldt by Friedrich Georg Weitsch, 1806

Travels and work in Europe

In 1794 Humboldt was admitted to the intimacy of the famous Weimar coterie, and contributed to the *Lebenskraft, oder der rhodische Genius*. In the summer of 1790 he paid a short visit to England in company with Forster. In geological and botanical tour of Switzerland and Italy he had obtained in the meantime official employment: appointed this service to the state was regarded by him as only an apprenticeship to the service of science, he fulfilled its duties with the highest post in his department, but he was also entrusted with several important diplomatic missions. The death of his father, and severing his official connections, he waited for an opportunity to fulfil his long-cherished dream of travel.

Latin American expedition

On the postponement of Captain Baudin's proposed voyage of circumnavigation, which he had been officially invited to accompany, Humboldt left Paris for Marseille, and then to Naples, the designated botanist of the frustrated expedition, hoping to join Napoleon Bonaparte in Egypt. Means of transport, however, were not forthcoming, and the two travellers

convinced them to make Spanish America the scene of their explorations.

Armed with powerful recommendations, they sailed in the *Pizarro* from A Coruña on June 5, 1799, and arrived six days on the island of Tenerife. Humboldt and Bonpland then sailed for Cumaná, Venezuela on July 16. Humboldt visited the mission at Carime where he found the oil-bird, which is now known as *Steatornis caripensis*. Returning to Cumaná, Humboldt observed, on the night of November 11–12, a brilliant meteor shower (the Leonids). He proceeded with Bonpland to Caracas in February 1800, on board with the purpose of exploring the course of the Orinoco River. This trip, which lasted four months, and covered 1,725 miles (2,776 km) of wild and largely uninhabited country, had the important result of establishing the existence of the Casiquiare canal (a communication between the water-systems of the rivers Orinoco and Amazon), and of determining the exact position of the bifurcation, as well as documenting the life of several native tribes such as the Maipures and their extinct rivals the Atures. Arriving in Bogotá on March 19, 1800, Humboldt and Bonpland discovered and captured some electric eels. They both received potentially dangerous electric shocks during their investigations.

On November 24, 1800, the expedition set sail for Cuba, and after a stay of some months they reached the main port of Cartagena, Colombia, on January 6, 1802, and commenced a difficult journey. Their first

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

next week

temporal information retrieval

today

temporal information extraction

maybe later

geographic and event-centric information retrieval

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

Greece Makes 'Good Progress' in Payment Talks

By Maria Petrakou and Nikolaos
Sep 20, 2011 10:39 PM GMT+0200 [\[add to rss feed\]](#)

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

The telephone talks tonight, which took place yesterday, were intended to damp concerns that Greece would miss its reduction targets and to clear the way for a sixth instalment of eurozone funds. The EU statement said a "full mission" will return to Athens next week. 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 are determined to tackle once and for all."

The EU said it expects Greece to make a payment of 1 billion euros in the next month. Finance Minister George Papandreou fights

1979 Soviet invasion

October 31, 1979 marks the day when the Afghan Armed Forces were overthrown by the Soviet Union. The invasion was a major event in the history of the Soviet Union and the Middle East. It led to the Soviet-Afghan War, which lasted until 1989. The invasion was a major event in the history of the Soviet Union and the Middle East. It led to the Soviet-Afghan War, which lasted until 1989.

October 31, 1979 marks the day when the Afghan Armed Forces were overthrown by the Soviet Union. The invasion was a major event in the history of the Soviet Union and the Middle East. It led to the Soviet-Afghan War, which lasted until 1989.

October 31, 1979 marks the day when the Afghan Armed Forces were overthrown by the Soviet Union. The invasion was a major event in the history of the Soviet Union and the Middle East. It led to the Soviet-Afghan War, which lasted until 1989.



The Soviet invasion



Travels and work in Europe
In 1794 he was a member of the Prussian Academy of Sciences and contributed to the Prussian periodical, Die Horen, a philosophical essay on the Lebenskraft, oder der menschlichen Seele. In 1795 he made a short visit to Copenhagen. In 1796 he made a general tour through Switzerland and Italy. He had obtained in the meantime official employment by appointment as professor of mines at Berlin. February 29, 1792, to the state was regarded by him as only an apprenticeship to the service of science, he fulfilled his duties with such conspicuous ability that not only did he rise rapidly to the highest post in his department, but he was also entrusted with several important diplomatic missions. The death of his mother, November 19, 1796, was the best of his genius, and securing no closer connections, he waited for an opportunity to fulfil his long-cherished dream of travel.

News articles.

Narrative documents.

Biographies.

Is time that important?

temporal information

has important key characteristics

Temporal information is **well-defined**:

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

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

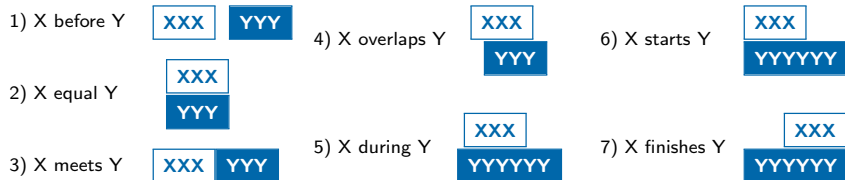
Is time that important?

temporal information

has important key characteristics

Temporal information is **well-defined**:

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



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**

TimeML **TIMEX3** tags,
value attribute

YYYY-MM-DD"T"HH:mm

e.g., 2016-06-16T14:33

→ Temporal information is **term- and language-independent**

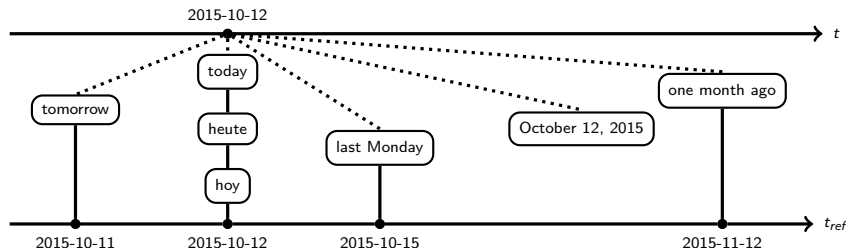
Is time that important?

temporal information

has important key characteristics

Temporal information can be **normalized**:

- expressions with **same semantics** → **same value**



Source: [Strötgen & Gertz 2016]

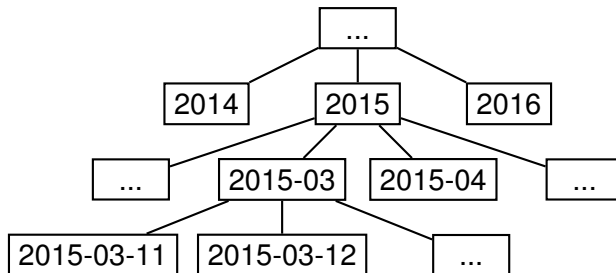
Is time that important?

temporal information

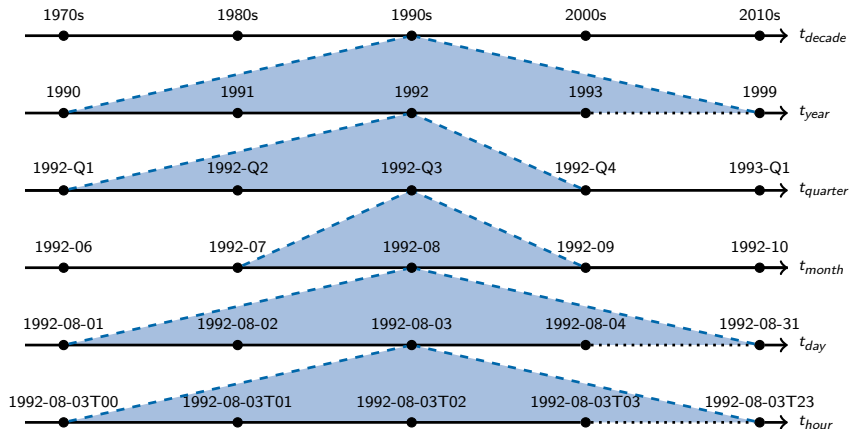
has important key characteristics

Temporal information can be **organized hierarchically**:

- expressions of different granularities



Is time that important?



Source: [Strötgen & Gertz 2016]

points in time on timelines of different granularities

Temporal Tagging

temporal expressions

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

temporal tagging

extraction and normalization of temporal expressions

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 Petrakis and Natalie

Sep 20, 2011 10:39 PM GMT+0200

ADD TO QUEUE

Recommend 10

Tweet 33

Share 4

2

Enlarge image



Evangelos Venizelos, Greece's finance minister, listens to an aide during a

Sept. 20, 2011
Taronis/Bloomberg

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

The telephone meeting **tonight**, which followed **yesterday**, were intended to damp 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 **next week** 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 will wedge that our financial data and economic growth are on track." The euro is determined to tackle once and for all.

The EU's next payment for **Greece** is on track to be released **next month** by Finance Minister **George Papandreu** fights

main challenge

ambiguities, e.g.,
may, march, fall

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** [ADD TO QUEUE](#)

[Recommend](#) 10

[Tweet](#) 33

[Share](#) 4

[+1](#) 2

[Enlarge image](#)



Evangelos Venizelos, Greece's finance minister, listens to an aide during a

Sept. 20, 2011
Taronis/Bloomberg

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

The telephone me**tonight**, ch follow **yesterday**, were intended to damp concerns that Greece 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 **next week** 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 str**earlier today** wledge that his financial data and econo**next month** determined to tackle once and for all.

The EL t payment for **Greece** is on track to be rele**next month** inister **George Papandreu** fights

tonight → **2011-09-20TNI**

yesterday → **2011-09-19**

next week → **2011-W39**

Sept. 20, 2011 → **2011-09-20**

next month → **2011-10**

main challenge

normalization of relative and underspecified expressions

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

different realizations of 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

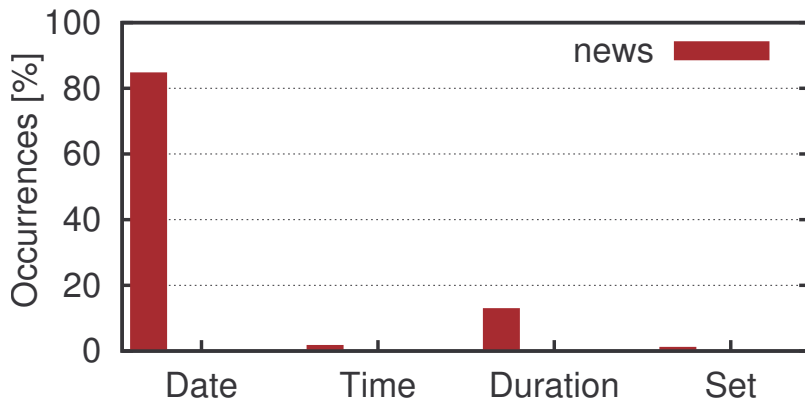
... On Thursday, the Census Bureau will publish the official population count for the United States, including the state-by-state totals required under the Constitution to determine how many seats each state is allocated in the House. The figures, eagerly awaited by many state government officials, are 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]

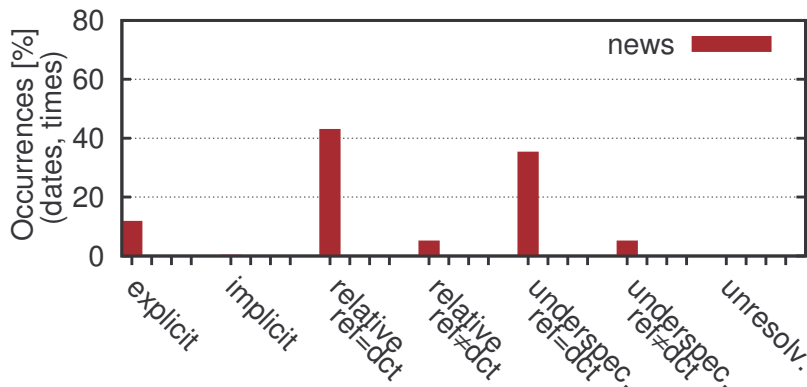
news articles with manually annotated temporal expressions



Temporal expressions in various corpora

TimeBank corpus [Pustejovsky et al. 2003]

news articles with manually annotated temporal expressions



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

Document Creation Time: 2009-12-19

...
1979: Soviet invasion

On December 7, 1979, Soviet informants to the Afghan Armed Forces ...and began to land in Kabul on December 25.

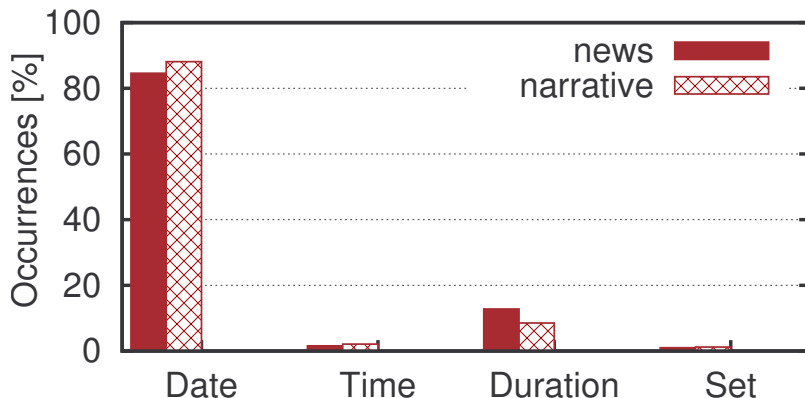
On December 27, 1979, 700 Soviet troops dressed in Afghan uniforms, ... That operation began at 19:00 hr., ... The operation was fully complete by the morning of December 28, 1979. ... According

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]

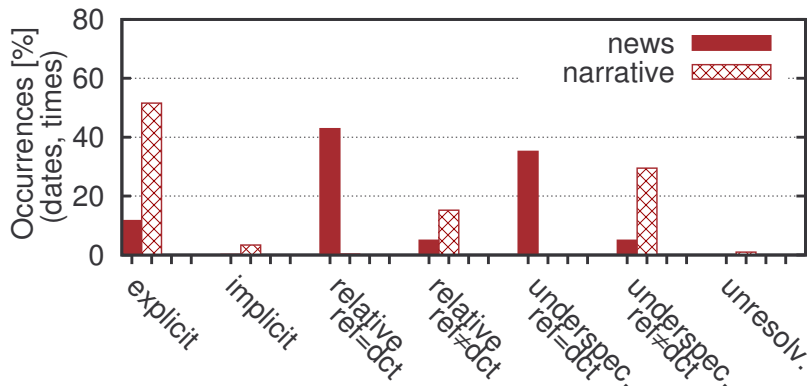
Wikipedia articles with manually annotated temporal expressions



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

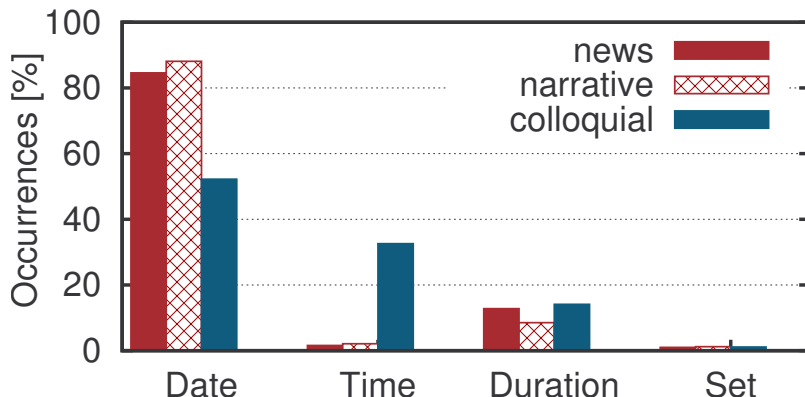
SMS	2010-01-10T05:19
Whats it u wanted 2 say last nite ?	
SMS	2010-09-23T09:50
Yo! Rem to come for lab tmr :-)	
SMS	2011-02-16T12:42
... andy is available at 10 am in his office	

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

Temporal expressions in various corpora

Time4SMS corpus [Strötgen & Gertz 2012]

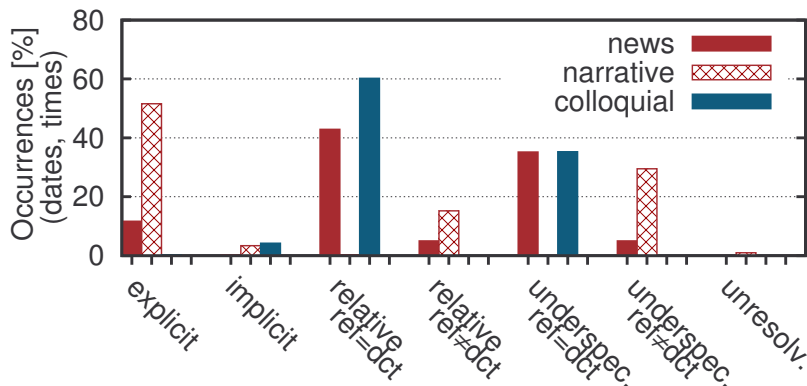
short messages with manually annotated temporal expressions



Temporal expressions in various corpora

Time4SMS corpus [Strötgen & Gertz 2012]

short messages with manually annotated temporal expressions



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

Scientific 2009-12-19

... Subjects consumed one tablet per day containing

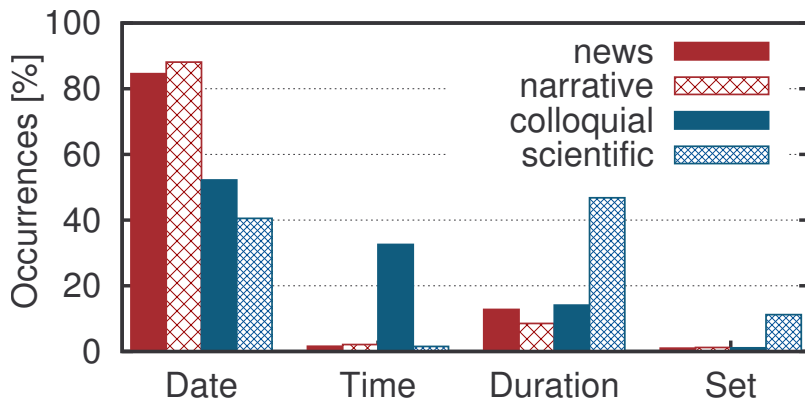
... Subjects were assessed at baseline, three and six months ... Clinical pathology analysis was performed at baseline and six months later ...

- often no real reference time
- local semantics (document time frame)
- “time point zero”

Temporal expressions in various corpora

Time4SCI corpus [Strötgen & Gertz 2012]

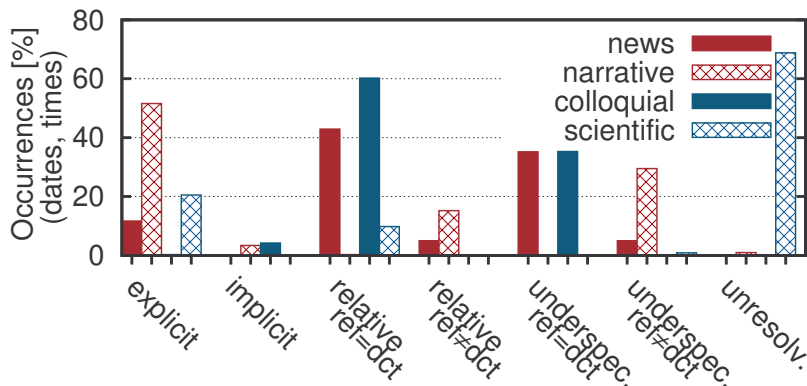
clinical trials with manually annotated temporal expressions



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 OO O O O O O B I IO

Machine learning-based extraction

frequently used classifier

- maximum entropy
- support vector machines
- conditional random fields

in the tutorial

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 $\rightarrow$ 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

system prediction	gold standard (ground truth)	
	positive	negative
positive	TP	FP
negative	FN	TN

Evaluation

system prediction	gold standard (ground truth)	
	positive	negative
positive	TP	FP
negative	FN	TN

measures

$$p = \frac{TP}{TP+FP}$$

$$r = \frac{TP}{TP+FN}$$

$$f_1 = \frac{2 \cdot p \cdot r}{p+r}$$

Evaluation

extraction only!

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

Evaluation

most meaningful

value f_1 relaxed matching

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

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

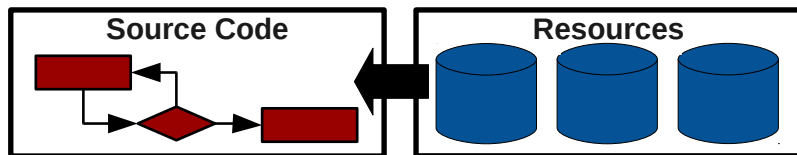
Extraction

- 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

// Pattern	// Pattern	// Pattern
// resource	// resource	// resource
// for month	// for month	// for month
// names (long)	// names (short)	// (number)
// access using:	// access using:	// access using:
// reMonthLong	// reMonthShort	// reMonthNum
January	Janl.?	10
February	Feb.?	11
March	Mar.?	12
April	Apr.?	0?[1-9]
...	...	

Normalization files:

- contain normalized values of such terms

```
// Normalization resource
// month names, numbers
// access using:
// "normMonth"
"January", "01"
"Jan.", "01"
"Jan", "01"
"01", "01"
"1", "01"
"February", "02"
...
```

HeidelTime's Language Resources

After read by resource interpreter, accessible by rules:

```
reMonthLong = (January|February|...)  
reMonthShort = (Jan\.?|Feb\.?|Mar\.?|...)  
reMonthNumber = (10|11|12|0?[1-9])  
...
```

```
normMonth("January") = "01"  
normMonth("Jan.") = "01"  
normMonth("Jan") = "01"  
normMonth("01") = "01"  
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**
 group(1) group(2) group(3)

NORM_VALUE="group(3)-%normMonth(group(1))-%normDay(group(2))"
 ="2010-%normMonth(**January**)-%normDay(**8**)"
 ="2010-01-08"

Simple rule example

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)**

[Moriceau & Tannier 2014, Camp & Christiansen 2012, Skukan et al. 2014]

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]

corpus	strategy	extraction	normalization
news	news	91.1	78.6
	narratives	91.1	61.5
narrative	news	87.9	56.9
	narratives	87.9	78.7

Don't trust a tagger developed for news if you want to process narratives (e.g., Wikipedia)

HeidelTime is Publicly Available

that Labor Day can always be calculated for the specific event specified as an implicit temporal expression, e.g., "the 2010", which took place on July 11, 2010. The HeidelTime temporal tagger is extensible with respect to the manner. For this, it is important that a temporal expression cannot be normalized without further context information. Expressions are marked with solid boxes. For example, "the following year" but also for underspecified expressions, e.g., "December 25", a reference time has to be specified. This reference time can either be the document time (for "today"), or another temporal expression in the document, e.g., "the following year". Sometimes, also the temporal relationship has to be specified. For instance, in Fig. 1a, "December" without knowing the relationship to the reference time. The tense of the sentence can be used to determine this relationship: present and future tense refer to a upcoming point in time, past tense refers to a previous point in time. In the example, "cited"), and thus "December" is normalized to "1997".

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

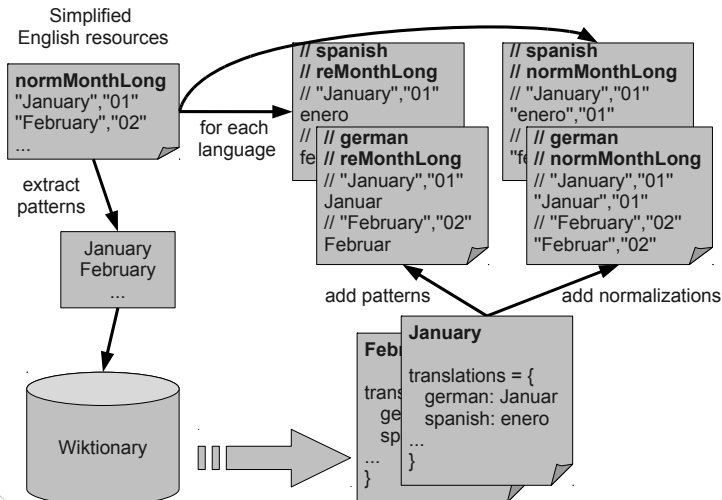
Developing Language Resources Automatically

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”



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

	Heidelberg 1.9 (manual)					Heidelberg – automatic				
	relaxed extr			value		relaxed extr			value	
language: corpus	P	R	F1	F1	acc.	P	R	F1	F1	acc.
ar: Arabic test-50*	90.9	90.9	90.9	82.2	90.4	91.7	31.8	47.2	38.0	80.5
ch: TE-2 test impro.	95.8	89.3	92.4	79.5	86.0	100	9.5	17.3	7.6	44.0
hr: WikiWarsHR	92.6	90.5	91.5	80.8	88.3	87.3	6.8	12.6	9.7	77.0
fr: FR-TimeBank	91.9	90.1	91.0	73.6	80.9	87.2	59.5	70.8	54.6	77.1
de: WikiWarsDE	98.7	89.3	93.8	83.0	88.5	98.4	64.7	78.1	59.7	76.4
it: EVALITA'14 test	92.7	86.1	89.3	75.0	84.0	98.5	41.2	58.1	49.3	84.9
es: TempEval-3 test	96.0	84.9	90.1	85.3	94.7	95.5	53.8	68.8	58.5	85.0
vn: WikiWarsVN	98.2	98.2	98.2	91.4	93.1	84.0	45.5	59.0	27.1	45.9
pt: PT-TimeBank test	87.3	75.9	81.2	63.5	78.2	91.5	59.3	72.0	59.4	82.5
pt: PT-TimeBank train	83.3	73.1	77.9	54.5	70.0	88.2	51.0	64.6	50.4	78.0
ro: Ro-TimeBank	–	–	–	–	–	31.9	11.4	16.9	7.8	46.2

Evaluation

On publicly available corpora

- compare automatically generated resources with Heidelberg's manually created resources
- worse, but very promising results (for many languages)

Before

- Heidelberg supported 13 languages
and no other temporal taggers for other languages available

Heidelberg 2.0:

- Heidelberg 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

several
pre-processing
components in
Elasticsearch

- 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

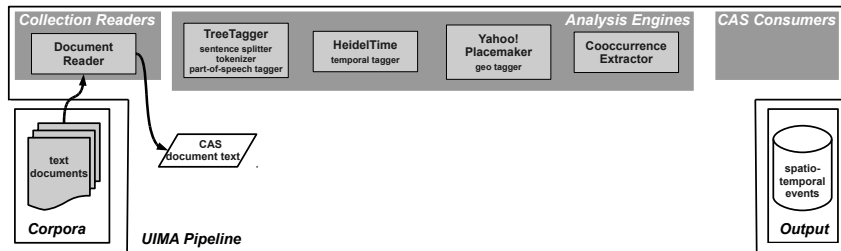
The Pipeline Principle – Why a (UIMA) Pipeline



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

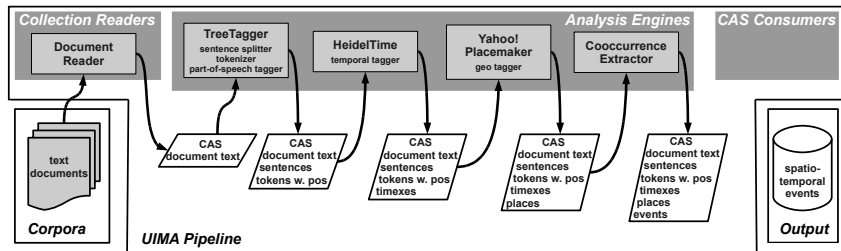
The Pipeline Principle – Why a (UIMA) Pipeline



Analysis Engines

- usually several analysis engines

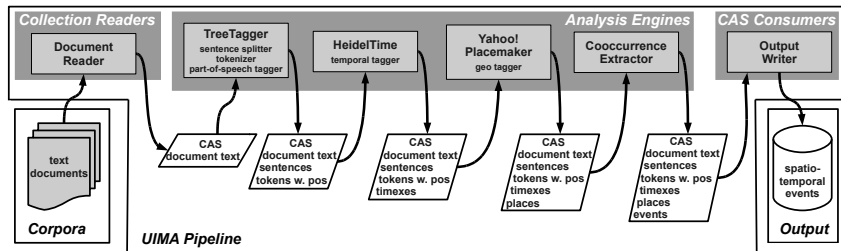
The Pipeline Principle – Why a (UIMA) Pipeline



Analysis Engines

- read the CAS
- analyze the documents (document text)
- add annotations to the CAS

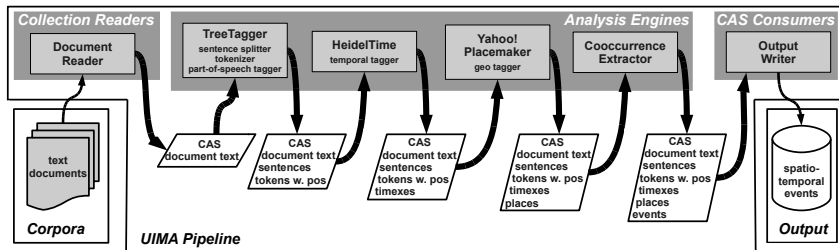
The Pipeline Principle – Why a (UIMA) Pipeline



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:

- Strötgen & Gertz (2016): Domain-sensitive Temporal Tagging, Morgan & Claypool Publishers (to appear).

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.
- Strötgen & Gertz 2016: Domain-sensitive Temporal Tagging (M&CP, to appear).
- Pustejovsky et al. 2005: Temporal and Event Information in Natural Language Text, LRE journal.
- Pustejovsky et al. 2003: The TimeBank Corpus, Corpus Linguistics.
- UzZaman et al. 2013: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations, SemEval.
- 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.
- Chang & Manning 2013: SUTime: Evaluation in TempEval-3, SemEval.
- Strötgen & Gertz 2010: HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions, SemEval.
- Strötgen & Gertz 2013: Multilingual and Cross-domain Temporal Tagging, LRE journal.
- Strötgen et al. 2013: HeidelTime: Tuning English and Developing Spanish Resources for TempEval-3, SemEval.
- Bethard 2013: ClearTK-TimeML: A Minimalist Approach to TempEval 2013, SemEval.**
- Verhagen et al. 2010: SemEval-2010 Task 13: TempEval-2, SemEval.
- 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.
- Skukan et al. 2014: HeidelTime.Hr: Extracting and Normalizing Temporal Expressions in Croatian, LTC.
- 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.
- Kuzey et al. 2016b: Temponym Tagging: Temporal Scopes for Textual Phrases, TempWeb.