Finding Salient Dates for Building Thematic Timelines

Rémy Kessler [1]
Xavier Tannier [1, 2]
Caroline Hagège [3]
Véronique Moriceau [1, 2]
André Bittar [3]

[1] Univ. Paris-Sud, France

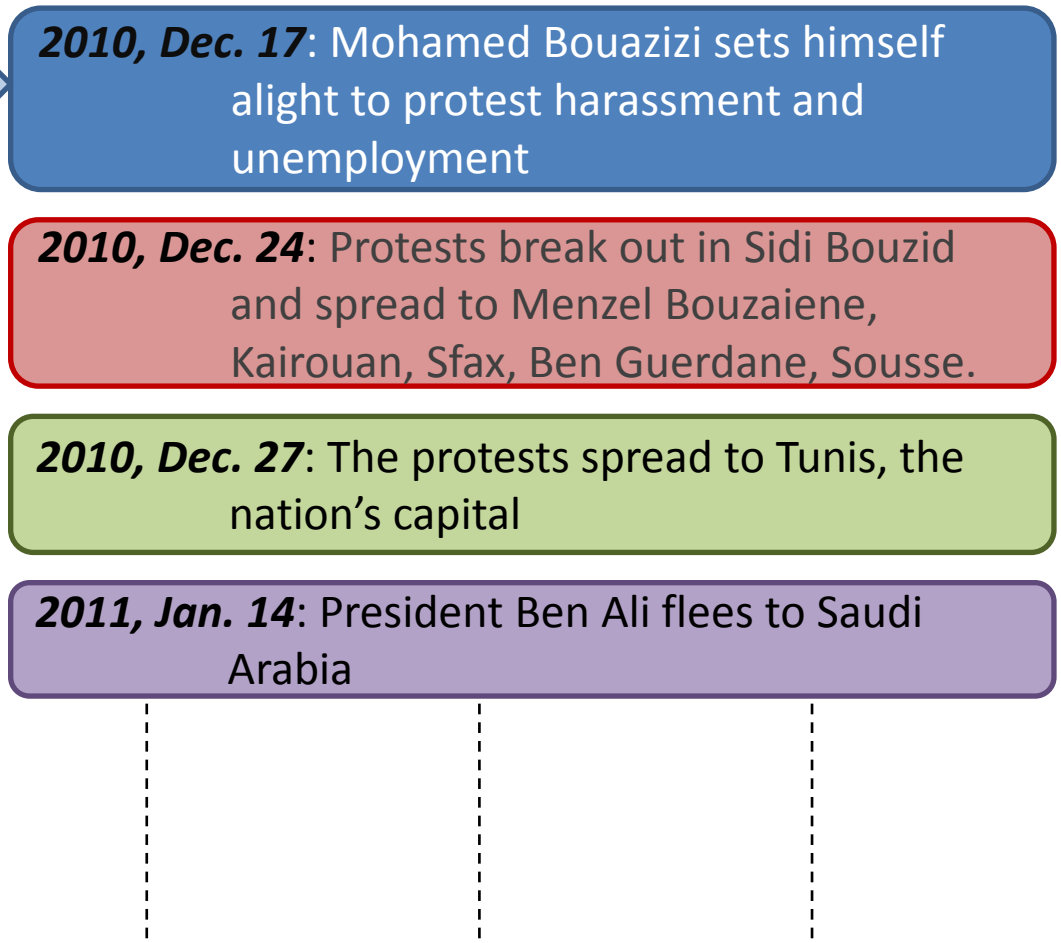
[2] LIMSI-CNRS, France

[3] Xerox Research Centre Europe, France

ACL 2012 Jeju, Republic of Korea

Context

- Our ultimate goal: build **automatic timelines** from a **query**
  - "Tunisian revolution"
  - "Michael Jackson"
  - "Wikileaks"
  - "accidents in Chinese mines"
  - etc.



Context

- Systems aiming at building timelines mainly see the problem as a traditional (multi-)document **summarization**

  See *Retrospective Event Detection*, *New Event Detection* (TDT) and others

  – Use of textual information (bag-of-words)

  – Use only **little temporal information**
    - Document creation time (DCT)
    - Rarely, absolute, full dates (day + month + year → 10th of July, 2012)

Only 7% of all dates (in our newswire corpus)

However, temporal information is crucial in timelines!

Our objectives

- Our ultimate goal: find **important events** from a **query**

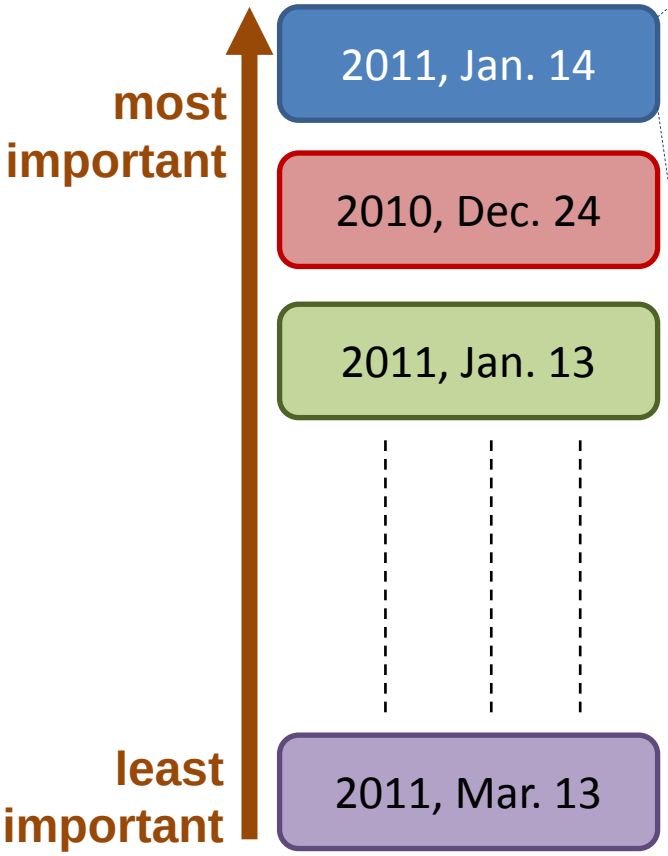
- Our intermediate goal (presented here): use temporal information to find **important dates**

  (Our assumption: important dates will lead to important events)

- Our system:
  - Extracts a maximum of temporal information from texts
  - Uses this information to extract ***salient dates***
  - Textual content is used only for the initial thematic document retrieval (*wrt* a query)

Our objectives

- Input: *"Tunisian revolution"* from: *2010* to: *now*
- Output:



Egypt's president Hosni Mubarak, who resigned on Friday, and *Tunisian president Zine El Abidine Ben Ali, who departed on January 14*, both bowed to unprecedented waves of popular protests.

The comments came after a Tunisian revolt which ended the 23 year old rule of *Ben Ali, who fled Tunisia for Saudi Arabia last Friday*.

Ben Ali signed his resignation on Friday after a wave of protests sparked by the suicide of a 26-year-old university graduate who was prevented by police from selling fruit and vegetables to make a living.

Resources

Corpus

- **AFP** (French news agency)
- 1.3 million texts in English, 2004-2011
- 511 documents/day
  (Lot of redundancy)
- 426 millions words
- XML file
  - Title
  - Document Creation Time (DCT)
  - Keywords
  - Textual content

```
<NewsML>
  …
<DateId>20110117T125527Z</DateId>
  …
  <HeadLine>Mauritanian sets himself on fire in govt protest: witness</HeadLine>
  …
  <body.content>
    <p>A Mauritanian set himself on fire in an anti-government protest Monday, witnesses said, […] </p>
    <p>Yacoub Ould Dahoud, 42, stopped his car in front of the Senate […] </p>
  </body.content>
</NewsML>
```

Reference "Chronologies"

- Textual event timelines
- Specific articles written by journalists in order to contextualize events.

```
<NewsML>
  …
  <DateId>20110114T142534Z</DateId>
  …
  <HeadLine>Timeline of Tunisian revolution</HeadLine>
  …
  <body.content>
    <p>President Ben Ali fled Friday to Saudi A
led to […]</p>
    <p>DECEMBER</p>
    <p> – 17 – </p>
    <p> Mohamed Bouazizi sets himself alight
unemployment</p>
…
```

Our aim:

- given a query,
- produce a list of dates
- where the dates of these reference chronologies are top ranked

Building Timelines

To extract salient dates, we need:

1. To get as many temporal information as possible

2. To define date salience:
   a) As a pure redundancy of information
   b) With linguistic filtering
   c) With other features (and with machine learning)

To extract salient dates, we need

“As many temporal information as possible”

(temporal and linguistic processing)

XIP

- XIP is a syntactic parser implemented at Xerox Research Centre Europe
- Deep grammatical dependency analysis
- Temporal expression recognition

XIP, temporal analysis

- **Precision-oriented date normalization**
  - Absolute dates
    - "January 5th, 2008"
    - 7% of the dates in the corpus (845,000)
  - DCT-related dates
    - "last Friday" or "on Friday" (use verb tense) ➔ `July 6th, 2012`
    - 40% of the dates in the corpus (4.6 millions)
  - No anaphoric dates ("the previous Friday")

- **Modality and Reported Speech information**
  - Temporal expressions linked to:
    - Future verbs
    - Modal verbs
    - Declaration verbs
    - Reported Speech

Building Timelines

To extract salient dates, we need:

1. To get as many temporal information as possible

2. To define date salience:
   a) As a pure redundancy of information
   b) With linguistic filtering
   c) With other features (and with machine learning)

Architecture

Temporal Analysis *(XIP)*

Indexing *(Lucene)*

INDEX

Offline

Querying

Filtering

Ranking Dates

27

7

22

Online (query-based)

Building Timelines

To extract salient dates, we need:

1. To get as many temporal information as possible

2. To define date salience:
   a) As a pure redundancy of information
   b) With linguistic filtering
   c) With other features (and with machine learning)

To extract salient dates, we need to

“Define date salience as a pure redundancy of information”

(temporal and linguistic processing)

Document retrieval and date scoring

- Document retrieval
  - Indexing and search using Lucene at sentence-level
  - Given a query, retrieve top 10,000 sentences

- Date scoring

An adaptation of classical *tf.idf* for dates:

$$\boldsymbol{tf.idf(d)} = f(d).\log(\frac{N}{df(d)})$$

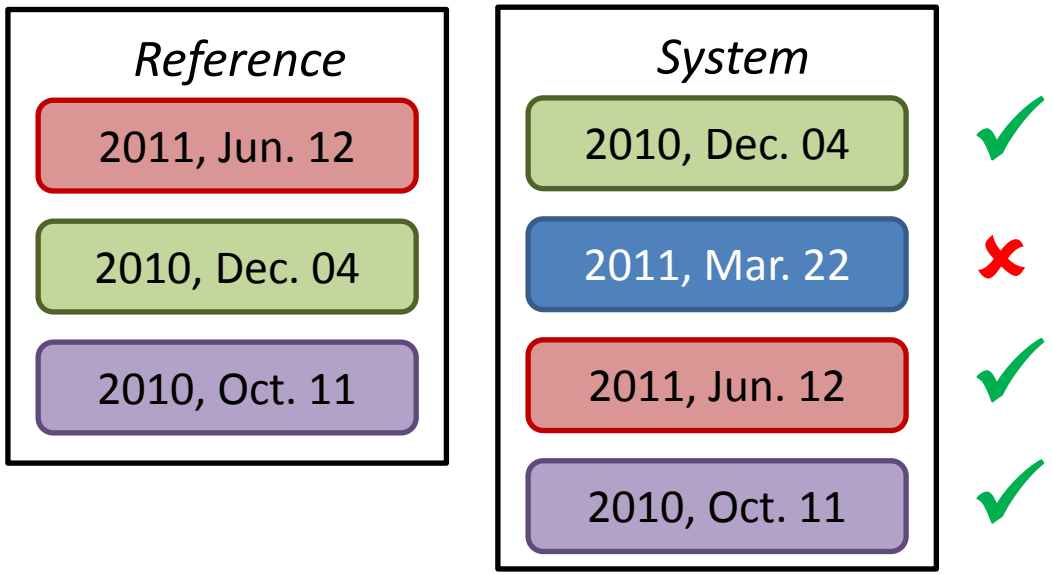
With

- $f(d)$ the number of occurrences of date d in the 10,000 sentences
- N the number of indexed sentences
- $df(d)$ the number of sentences containing d in the entire corpus

Evaluation

- What do we evaluate?
  - *Are dates from reference chronologies on the top of our ranked list of dates?*
  - (but the reference is subjective, we'll talk about this later)
  - No evaluation on associated text

- How do we evaluate?
  - Mean Average Precision (MAP)
  - On **91** manual chronologies from AFP corpus



Baselines

1. BL_{DCT}
   - Top 10,000 sentences
   - Only DCTs are considered
   - Dates are ranked by their $tf.idf(d)$

2. BL_{abs}
   - Top 10,000 sentences
   - Only absolute dates are considered
   - Dates are ranked by their $tf.idf(d)$

3. BL_{mix}
   - Absolute dates are considered
   - DCT when no absolute date in the sentence

Baseline Results

Baseline "only DCT"	
Model	**BL_{DCT}**
MAP score	**0.5523**
Baseline "only absolute dates"	
Model	BL_{abs}
MAP score	0.2778
Baseline "mixed"	
Model	BL_{mix}
MAP score	0.4135

Using XIP date normalization

1. SD
   - All **absolute and normalized relative dates** are considered
   - Dates are ranked by their $tf.idf(d)$

Salient Dates Results

Salient date run with all dates	
SD	0.6982

Building Timelines

To extract salient dates, we need:

1. To get as many temporal information as possible

2. To define date salience:
   a) As a pure redundancy of information
   b) With linguistic filtering
   c) With other features (and with machine learning)

Using XIP date normalization and filtering

1. SD
   - All **absolute and normalized relative dates** are considered
   - Dates are ranked by their $tf.idf(d)$

2. SD_X
   - Modality, future verbs and reported speech indicate that the event might not be factual
   - Filtering these events is intended to reduce noise
   - Filtering is achieved by removing dates associated with:
     - A reported speech verb (X = R)
     - A modal verb (X = M)
     - A future verb (X = F)
     - A declaration verb (X = D)
   - Filters can be combined

Salient Dates Results

Salient date runs with all dates	
SD	0.6982
Salient date runs with filtering	
SD_R	0.6996
SD_F	0.6993 **
SD_M	0.7005 *
SD_D	0.7091 **
SD_{FMD}	0.7091 **
SD_{RFMD}	0.7146 **

* : significant, $p < 0.05$
** : highly significant, $p < 0.01$

wrt SD run

Building Timelines

To extract salient dates, we need:

1. To get as many temporal information as possible

2. To define date salience:
   a) As a pure redundancy of information
   b) With linguistic filtering
   c) With other features (and with machine learning)

To extract salient dates, we need to

“Define date salience with other features”

(machine learning)

Learning salience: features

1. The more a date is mentioned, the more important it is
   - Sum of Lucene scores for all sentences containing the date
   - Number of sentences containing the date
   - …

2. An important event is still written about, a long time after it occurs
   - Distance (in days) between the date and the most recent mention of this date
   - Distance between the date and the DCT of the article where it appears

3. Other features
   - Lucene's best ranking of the date
   - Number of times where the date is absolute in the texts
   - …

Learning date salience

- Classification between ***salient dates*** and ***non-salient dates***
  - Dates in AFP chronologies are salient, all others are not
    (but the reference is subjective, we'll talk about very soon)
  - Used IcsiBoost, implementation of adaptative boosting
    (Freund and Shapire, 1997)
- Our aim is not to classify dates, but to **rank** them.
- We therefore used the **predicted probability** $P(d)$ of being salient, returned by the classifier
- $P(d)$ is mixed with values $tfidf(d)$:

$$score(d) = P(d) \times tfidf(d)$$

Machine Learning Results

- Cross-validation on the 91 chronologies

Machine Learning run
ML
0.7918 **

** : highly significant, p << 0.01 wrt SD runs

Conclusion

Result Discussion

1. Using simple frequency counting and normalization improves baseline by far (MAP ≈ 0.7)

2. Adding linguistic filtering for reducing noise leads to a significant improvement (MAP + 0.02 at best)

3. Adding features and machine learning is even much better (MAP ≈ 0.8)

Result Discussion

- **Salient dates are not timelines**, work still needs to be done there
  - Finding central/representative sentence(s) within the set of sentences associated to a date
  - Using clustering techniques inside these sentences
  - Adding little semantic analysis for determining the importance of events
  - Mixing summarization techniques with salient dates extraction (any people interested?)

Perspectives

- Evaluation of timelines is a big issue
  1. **Lack of references**
     - Need good experts in many domains
  2. **Problem of definition** of the "important event"
     - Otherwise the task is too subjective
     - Difficult to define "guidelines"
- We have answer to problem 1.
  - Reference chronologies, in English, French and Arabic, with new ones every week
    - Made by journalists that have no idea of what we do
    - With the corresponding corpus
    - But still very subjective
- Interested into making timelines over this corpus?
  - The corpus is NOT free (AFP) but this CAN be discussed.

Thanks!

(This work has been partially funded under ANR project Chronolines)