Examining the State-of-the-Art in News Timeline Summarization

Demian Gholipour Ghalandari, Georgiana Ifrim
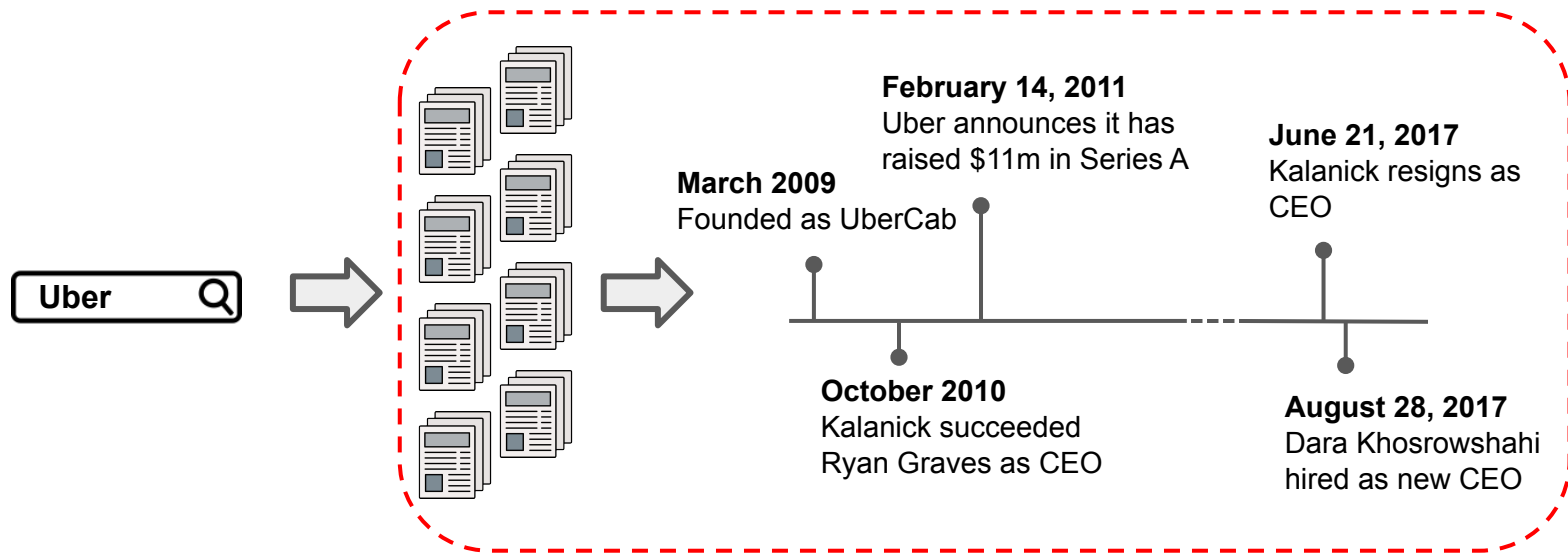
School of Computer Science
University College Dublin, Ireland

Outline

- Timeline Summarization (TLS) Task
- Previous Work
- Types of Approaches
  - Date-wise
  - Clustering-based
- New Dataset
- Experiments
- Conclusions

What is Timeline Summarization (TLS)?

March 2009
Founded as UberCab

February 14, 2011
Uber announces it has raised $11m in Series A

June 21, 2017
Kalanick resigns as CEO

October 2010
Kalanick succeeded Ryan Graves as CEO

August 28, 2017
Dara Khosrowshahi hired as new CEO

Uber 🔍

TASK

- **Input:** Article collection **A** + query (topic) keyphrases **Q**
- **Output:** List of <date, summary> pairs
- **Length constraints: L** dates, with **M** total sentences, or **K=M/L** sentences per date

Previous Work in TLS

Shortcomings in previous work
- Papers about isolated subtasks: date selection, date summarisation
- Inappropriate evaluation metrics, inconsistent experimental settings
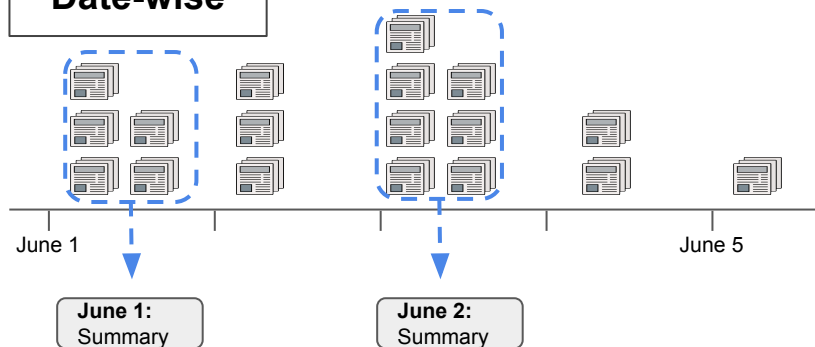- Small ground-truth datasets

Most recent SOTA (Martschat 2018')
- Establishes experimental settings, uses appropriate evaluation metrics
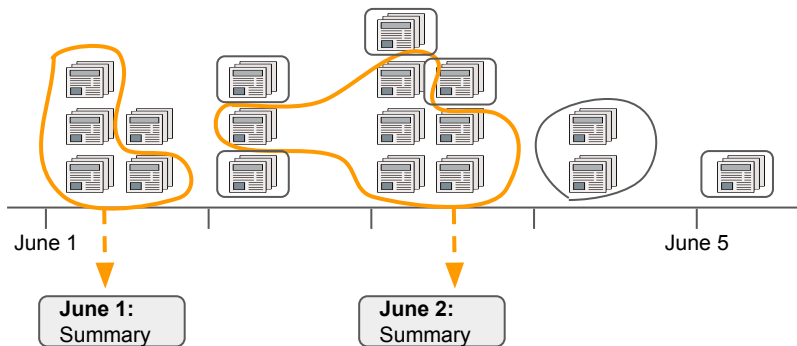- Method not scalable
- Poor comparison to previous approaches

➡️ Not clear how TLS can be approached and how well different strategies work
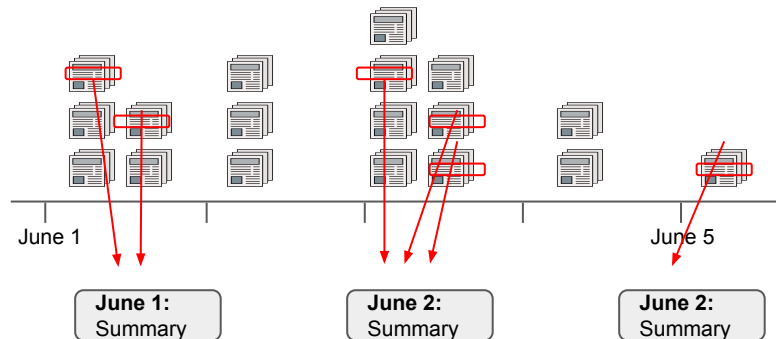
Approaches to Build News Timelines

Date-wise

June 1   June 5

June 1: Summary

June 2: Summary

Clustering / Event-based

June 1   June 5

June 1: Summary

June 2: Summary

One-stage summarization (current SOTA)

June 1   June 5

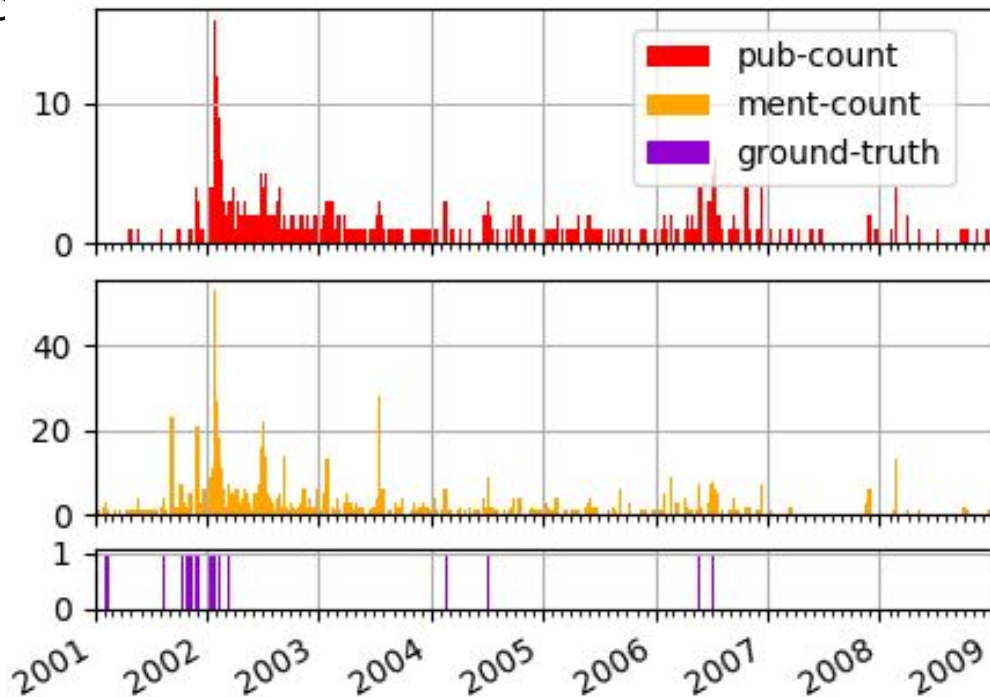**June 1:** Summary
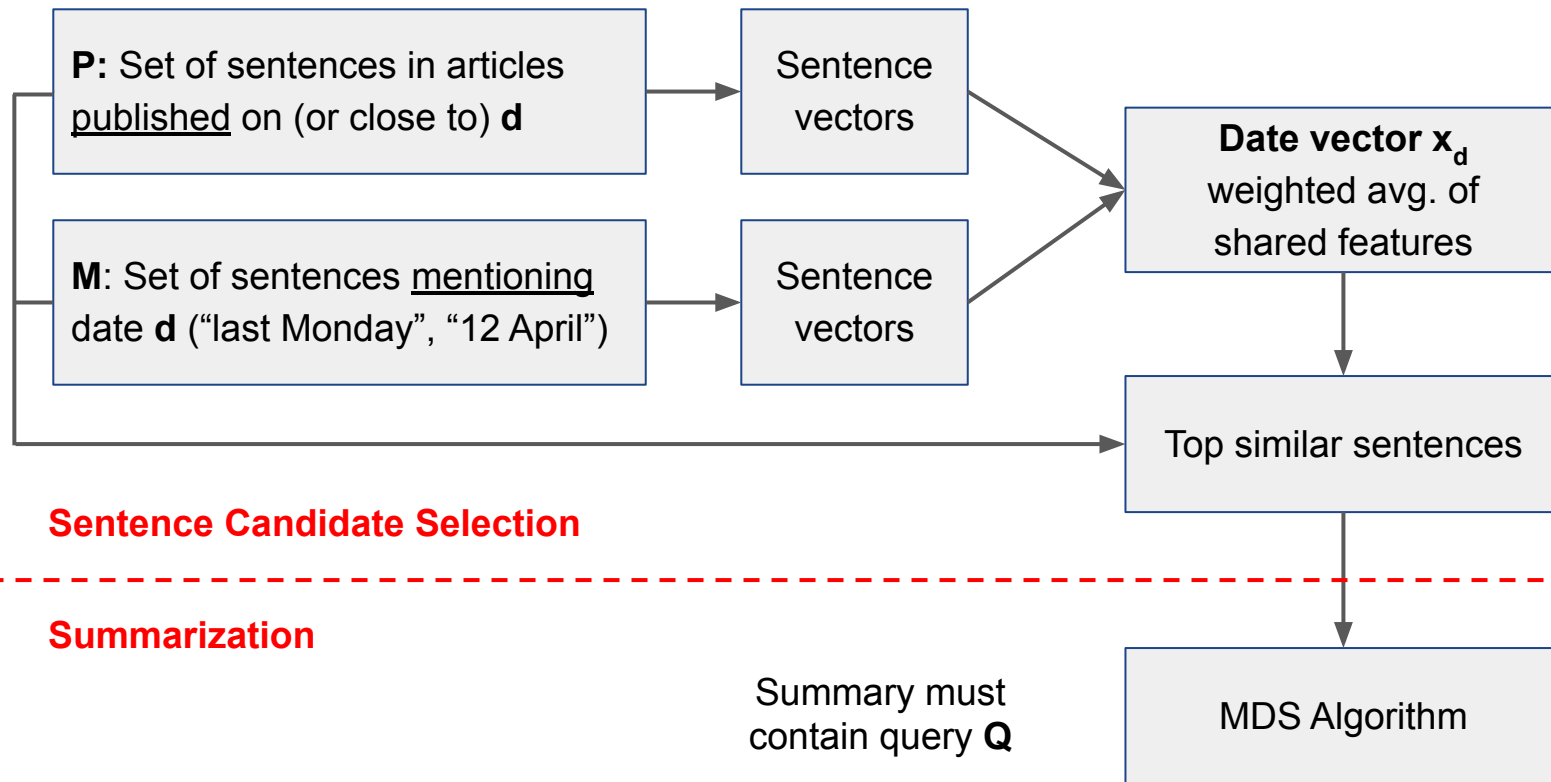
June 2: Summary

June 2: Summary

This work:
- Implement + evaluate date-wise & clustering-based TLS approaches
- Side-by-side comparison
- New & efficient SOTA by date-wise approach

Date-wise Approach (1/2): Date Selection

- Task: select **L** important dates that should
  appear on the timeline
- Regression with date features
  (Tran et al. 2013)
  - \# published articles
  - \# date is mentioned in text
    (e.g. 'yesterday', '2 May')
    - before/after/same day
- Alternative approaches
  - pub-count only
  - mention-count only
  - graph-based

Date-wise Approach (2/2): Summarizing a Particular Date

P: Set of sentences in articles <u>published</u> on (or close to) **d** → Sentence vectors

M: Set of sentences <u>mentioning</u> date **d** ("last Monday", "12 April") → Sentence vectors

Date vector x$_d$ weighted avg. of shared features

Top similar sentences

Sentence Candidate Selection

Summarization

Summary must contain query **Q**

MDS Algorithm

Clustering-based Approach

Select L important *events* (clusters), summarize these separately

Clustering
- Markov clustering on temporal article graph
- Nodes: articles
- Edges: cosine similarity between article vectors
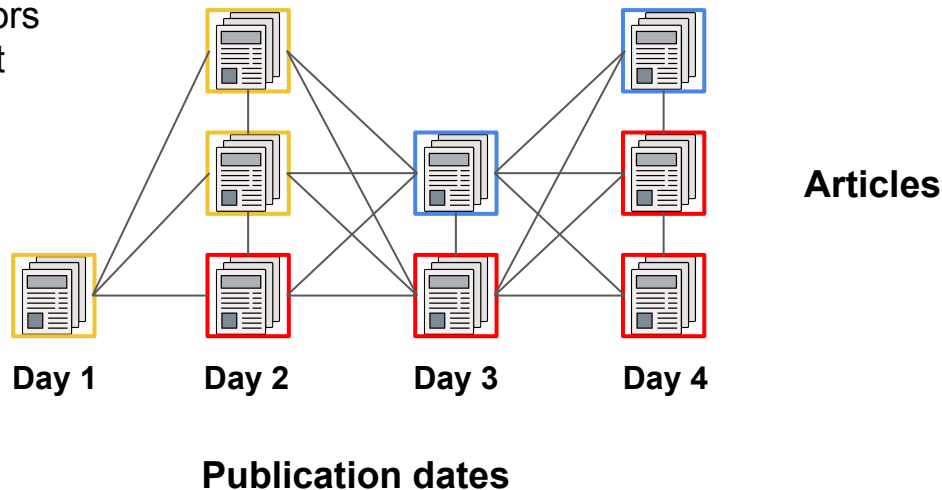- Edge only if publication dates are 1 day apart

Cluster dates
- Most frequently mentioned date in cluster

Cluster ranking: options
- Cluster size
- **Mention count of cluster date**
- Regression

Cluster summarization
- Unsupervised MDS algorithm

Articles

Day 1   Day 2   Day 3   Day 4

Publication dates

New Dataset for TLS ("Entities")

	#Topics	#TLs	Avg. articles	Avg. summary sentences	Avg. TL-dates	Avg. duration
T17	9	19	508	2.9	36	212 days
Crisis	4	22	2310	1.3	29	343 days
Entities	47	47	959	1.2	23	~12 years

- Larger dataset & longer time periods
- Ground-truth: timelines extracted from "CNN Fast Facts"
- Queries: timeline titles (e.g. "Steve Jobs")
- Input articles: from TheGuardian API
- Adjustments & Filtering: ensuring overlap between timeline & articles

Evaluation

- Metrics (Martschat '17)
  - Alignment-based ROUGE (AR) F-score
  - Date Selection F-score
- Leave-one-out cross-validation over individual tasks in a dataset

	T17 Dataset			Crisis Dataset			Entities Dataset		
	AR1-F	AR2-F	Date-F	AR1-F	AR2-F	Date-F	AR1-F	AR2-F	Date-F
One-stage (Martschat'18)	0.105	0.03	0.544	0.075	0.016	0.281	0.042	0.009	0.167
Date-wise	0.12	0.035	0.544	0.089	0.026	0.295	0.057	0.017	0.205
Clustering-based	0.082	0.02	0.407	0.061	0.013	0.226	0.051	0.015	0.174
Oracle	0.312	0.128	0.926	0.367	0.15	0.974	0.232	0.075	0.757

Observations & Discussion

Redundancy
- Common phenomenon: adjacent dates in automatic timelines (avoided by clustering-based TLS)
- Forcing date-wise approach to avoid redundancy decreases performance!

Clustering-based approach: often gets granularity wrong

Speed
- Date-wise & clustering-based much faster than submodular one-stage approach

When is the task particularly difficult?
- High compression ratio: large article collection, short timelines

Ground-truth
- Subjective in picking important events & formulating summaries (editorial decisions)

Conclusions & Future Outlook

- Compared 3 TLS strategies (date-wise, clustering, one-stage)
  - Date-wise approach obtains new SOTA
  - Date-wise & clustering very efficient compared to one-stage
- Textual date mentions are very helpful
- TLS is a hard task, difficult to imitate ground-truth
- Redundancy remains a problem
- Quality: reasonable apart from occasional redundancy & style issues

What can be improved?

- Better datasets: larger, more news sources, more timelines
- End-to-end methods, abstractive techniques
- Alternative automatic evaluation, human evaluation