AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UNIVERSITY OF SOUTH CAROLINA

CSCE 771: Computer Processing of Natural Language
Lecture 18: Event Extraction, ST reasoning

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

20TH OCTOBER, 2022

Carolinian Creed: "I will practice personal and academic integrity."

Organization of Lecture 18

- Opening Segment
  - Announcements: Poll Results
- Main Lecture

Main Section
- What is an event?
- Extraction and linking
- Spatio-temporal reasoning
- Applications

- Concluding Segment
  - About Next Lecture – Lecture 19

Recent Classes

Oct 11 (Tu)	Guest Lecture – Dr. Amitava Das: Using lang models to solve NLP tasks
Oct 13 (Th)	
Oct 18 (Tu)	Entity extraction, linking
Oct 20 (Th)	Events extraction, spatio-temporal analysis
Oct 25 (Tu)	Topic Analysis
Oct 27 (Th)	PROJ REVIEW
Nov 1 (Tu)	NLP Task: Sentiment
Nov 3 (Th)	NLP Task: Summarization

Review of Lecture 16

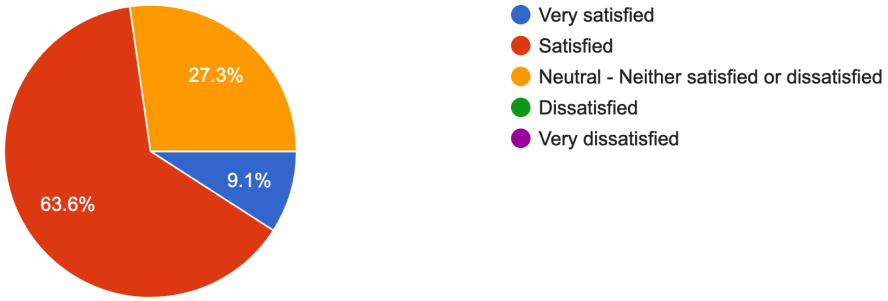
- Entity extraction
- Extraction linking
- Case study: can be done across text and figures; document sets

Announcements

- Mid-course survey results: https://forms.gle/JNMPvUMG2pvs1r5U9 (optional, due by Oct 20, 2022)

How satisfied are you with the course?

11 responses



Things to improve

- Clarify quiz format => in-class, programming, programming, in-class
- Explain how the computation works => Noted
- Explain project rubric => Next slide
- More project oriented discussion => Noted / Encourage 1:1
- Give format of (NLP) task by example first => Noted

Project Assessment Discussion

Choosing a Project – Some Considerations

- Scope: what is the problem?
- Current-state: what happens in the problem today?
- Who cares: who will benefit with the problem being solved?
- Desired-state: what will be the future situation if your project succeeds?
- Resources/ dataset: do you have reasonable data and compute resources to do the work?
- Evaluation: how will we measure goodness of the work?

From Class 5

Discussion: Course Project

- Expectations
  - Apply methods learned in class or of interest to a problem of interest
  - Be goal oriented: aim to finish, be proactive, be innovative
  - Do top-class work: code, writeup, presentation
- Typical pitfalls
  - Not detailing out the project, assuming data
  - Not spending enough time
- What will be awarded
  - Results and efforts (balance)
  - Challenge level of problem

From Class 5

Course Project – Deadlines and Penalty Rubric

- Project plan **not** ready by Sep 15, 2020 **[-20%]**
  - * Project Title
  - * Description: motivation and expected output
  - * Illustrative Test cases: i.e., Example input / output
  - * Data sources:
  - * Technique and tools to use:
  - * Metric for measuring output
  - * How will you collect results
  - * Format of report, presentation
  - * Time schedule:
- Project report **not** ready by Nov 10, 2022 **[-20%]**
- Project presentations **not** ready by Nov 15, 2022 **[-10%]**

From Class 5

- W1 - Sep 26
- W2 – Oct 3
  - **Review presentation for class: 3 min each – Oct 4, 2022**
- W3 – Oct 10
- W4 – Oct 17
- W5 – Oct 24
  - **Review presentation for class: 3 min each – Oct 27, 2022**
- W6 – Oct 31
- W7 – Nov 7
- W8 – Nov 14
- W9 – Nov 21

Project Rubric

- **Project results** – 60%
  - Working system ? – 30%
  - Evaluation with results superior to baseline? – 20%
  - Considered related work? – 10%
- **Project efforts** – 40%
  - Project report – 20%
  - Project presentation (updates, final) – 20%

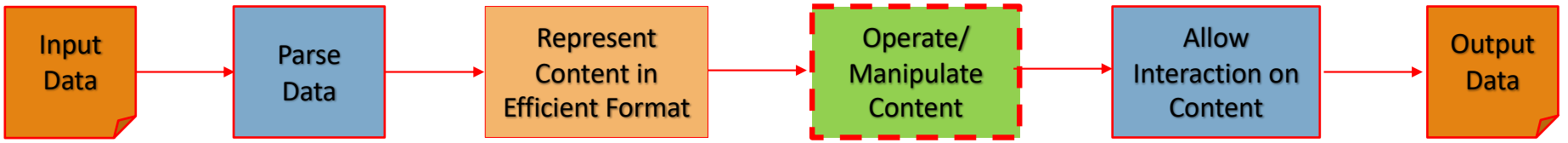
- **Bonus**
  - Challenge level of problem – 10%
  - Instructor discretion – 10%
- **Penalty**
  - Lack of timeliness as per announced policy (right) - up to 60%

Milestones

- Penalty: **not** ready by Sep 15, 2022 **[-20%]**
- Project report **not** ready by Nov 10, 2022 **[-20%]**
- Project presentations **not** ready by Nov 15, 2022 **[-10%]**

Main Lecture

Methods to Extract Content



Motivation for Events Extraction

- Event - A type of entity which has much practical interest
  - Business intelligence
  - Perceiving changes in the world
- There is rise in technologies for detection of events from text data sources
  - News
  - Social Media
  - Technical papers

IBM acquired Red Hat

Terminology and Representation

- Event model: <**label**, **time**>
  - Example: IBM acquired Red Hat
- Spatio-temporal event model: < **label**, **time**, **location**>
  - Example: IBM, headquartered at Armonk, NY, acquired Red Hat, which is based in … on Oct 28, 2018
- Spatio event: < **label**, **location**>
  - Example: IBM, headquartered at Armonk, NY, acquired Red Hat.
  - Time implicit

Different Refences to a Location

- Informal names – here, there, near
- Country names (ISO 3166 — Country Codes)
  - https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes
- States, County names
- City names
- Zip code
- Latitude /Longitude

ISO 3166[1]			ISO 3166-1[2]			ISO 3166-2[3]	
Country name[5]	Official state name[6]	Sovereignty[6][7][8]	Alpha-2 code[5]	Alpha-3 code[5]	Numeric code[5]	Subdivision code links[3]	Internet ccTLD[9]
Afghanistan	The Islamic Republic of Afghanistan	UN member state	AF	AFG	004	ISO 3166-2:AF	.af
Akrotiri and Dhekelia – See United Kingdom, The							
Åland Islands	Åland	Finland	AX	ALA	248	ISO 3166-2:AX	.ax
Albania	The Republic of Albania	UN member state	AL	ALB	008	ISO 3166-2:AL	.al
Algeria	The People's Democratic Republic of Algeria	UN member state	DZ	DZA	012	ISO 3166-2:DZ	.dz
American Samoa	The Territory of American Samoa	United States	AS	ASM	016	ISO 3166-2:AS	.as
Andorra	The Principality of Andorra	UN member state	AD	AND	020	ISO 3166-2:AD	.ad
Angola	The Republic of Angola	UN member state	AO	AGO	024	ISO 3166-2:AO	.ao
Anguilla	Anguilla	United Kingdom	AI	AIA	660	ISO 3166-2:AI	.ai
Antarctica[a]	All land and ice shelves south of the 60th parallel south	Antarctic Treaty	AQ	ATA	010	ISO 3166-2:AQ	.aq

Source: https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes

Different Refences to Time

- Exact v/s approximate references
  - 6 am
  - Ambiguous – soon, afterwards
- Different types
  - Absolute
  - Relative
  - Duration
- Language forms
- Standards/ Conventions
  - dd/mm/yyyy
  - ISO 8601 Data elements and interchange formats

Absolute	Relative	Durations
April 24, 1916	yesterday	four hours
The summer of '77	next semester	three weeks
10:15 AM	two weeks from yesterday	six days
The 3rd quarter of 2006	last quarter	the last three quarters

Category	Examples
Noun	*morning, noon, night, winter, dusk, dawn*
Proper Noun	*January, Monday, Ides, Easter, Rosh Hashana, Ramadan, Tet*
Adjective	*recent, past, annual, former*
Adverb	*hourly, daily, monthly, yearly*

Extraction Methods

- Entities
  - Extract entities using extraction methods (regex, lookup, learning based)
  - Link to authoritative entities – disambiguation

IBM acquired Red Hat

Big Blue acquired Red Hat

- Entities relationships
  - Use language properties to navigate
    - Parse and use dependency graphs
  - Use rules

IBM acquired Red Hat

Exercise – "Nobel Prize 2022"

Source 1:
https://www.nobelprize.org/prizes/lists/all-nobel-prizes/

Source 2:
https://www.weforum.org/agenda/2022/10/nobel-prize-winner-peace-medicine-physics/

Review Sources and Answer These Questions:

- Where are events (related to wins) easy to extract using entity extraction / linking methods?
- Are the two sources consistent with win information?

2022

The Nobel Prize in Physics 2022

Alain Aspect, John F. Clauser and Anton Zeilinger "for experiments with entangled photons, establishing the violation of Bell inequalities and pioneering quantum information science"

The Nobel Prize in Chemistry 2022

Carolyn R. Bertozzi, Morten Meldal and K. Barry Sharpless "for the development of click chemistry and bioorthogonal chemistry"

The Nobel Prize in Physiology or Medicine 2022

Svante Pääbo "for his discoveries concerning the genomes of extinct hominins and human evolution"

The Nobel Prize in Literature 2022

Annie Ernaux "for the courage and clinical acuity with which she uncovers the roots, estrangements and collective restraints of personal memory"

The Nobel Peace Prize 2022

Ales Bialiatski, Memorial, and Center for Civil Liberties. The Peace Prize laureates represent civil society in their home countries. They have for many years promoted the right to criticise power and protect the fundamental rights of citizens. They have made an outstanding effort to document war crimes, human right abuses and the abuse of power. Together they demonstrate the significance of civil society for peace and democracy.

Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2022

Ben S. Bernanke, Douglas W. Diamond and Philip H. Dybvig "for research on banks and financial crises".

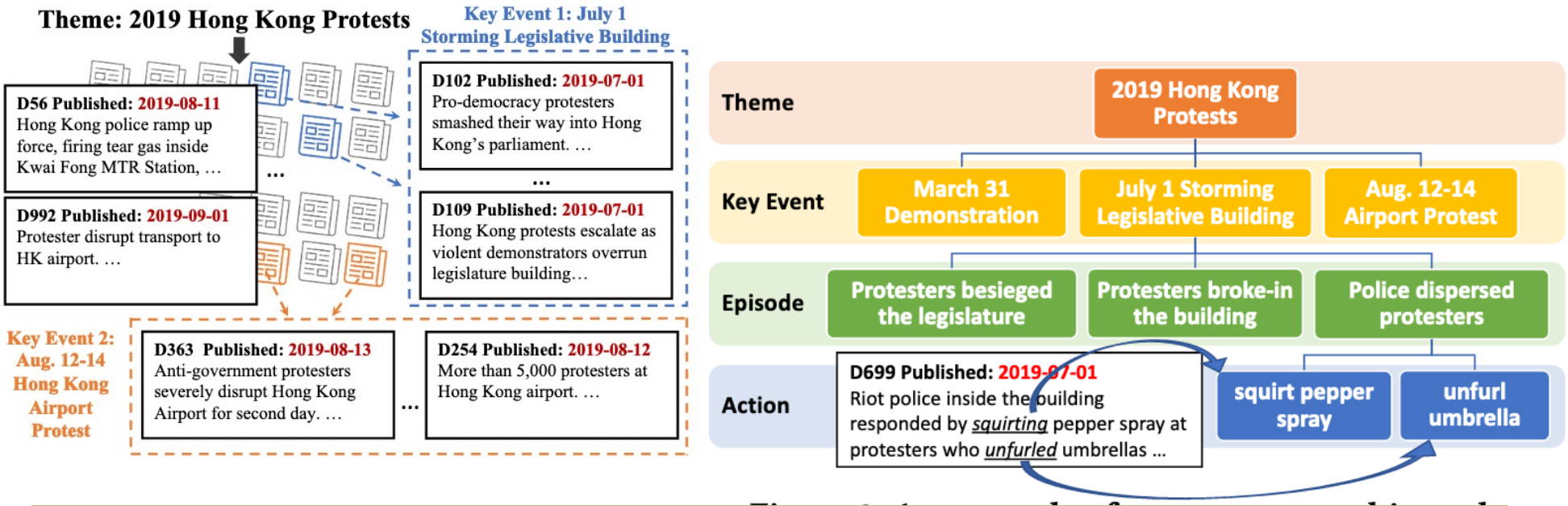
Credit: Source 1

“Levels” of Event from a Documents

Level	Data	Example	Comments
Theme level	Corpus	"2019 Hong Kong Protests”	Very similar to topic analysis
Key event level	Corpus	“HK Airport Protest on Aug. 12-14"	Multiple action events happening at proximal location and time
Action level	One document mention	"the police hit the left arm of the protester"	Implies a precise location and time

Unsupervised Key Event Detection from Massive Text Corpora
Yunyi Zhang, Fang Guo, Jiaming Shen, Jiawei Han, KDD 2022

Events from News at Various Levels



Figures source: Unsupervised Key Event Detection from Massive Text Corpora, Yunyi Zhang, Fang Guo, Jiaming Shen, Jiawei Han, KDD 2022

EvMine Approach Summary

Key Idea: "temporal term frequency–inverse time frequency" -- (ttf-itf) measure

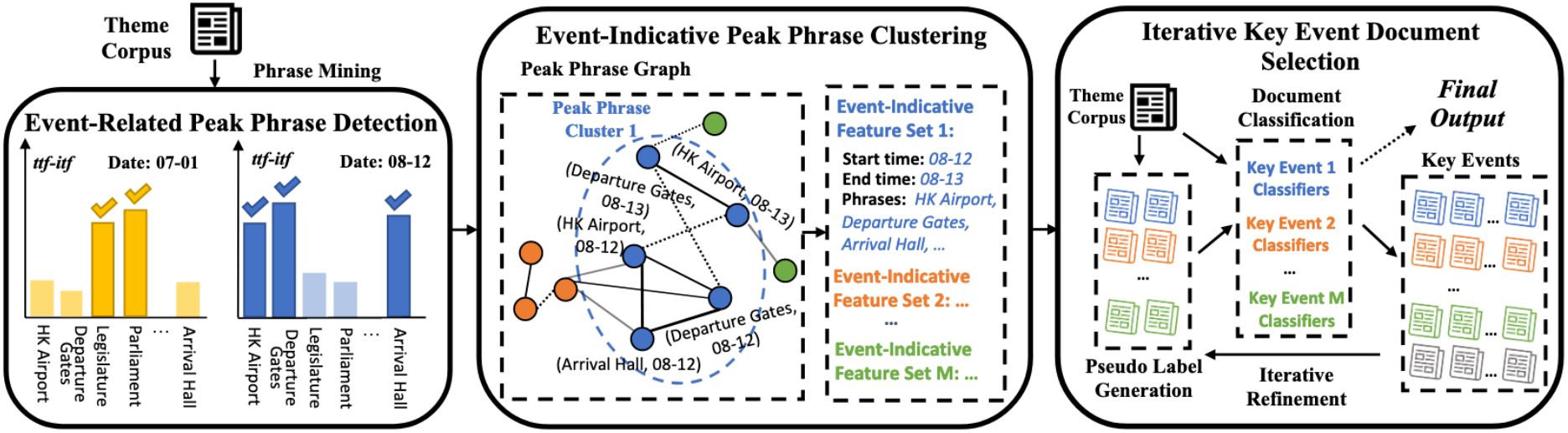


Figure source: Unsupervised Key Event Detection from Massive Text Corpora, Yunyi Zhang, Fang Guo, Jiaming Shen, Jiawei Han, KDD 2022

EvMine Output

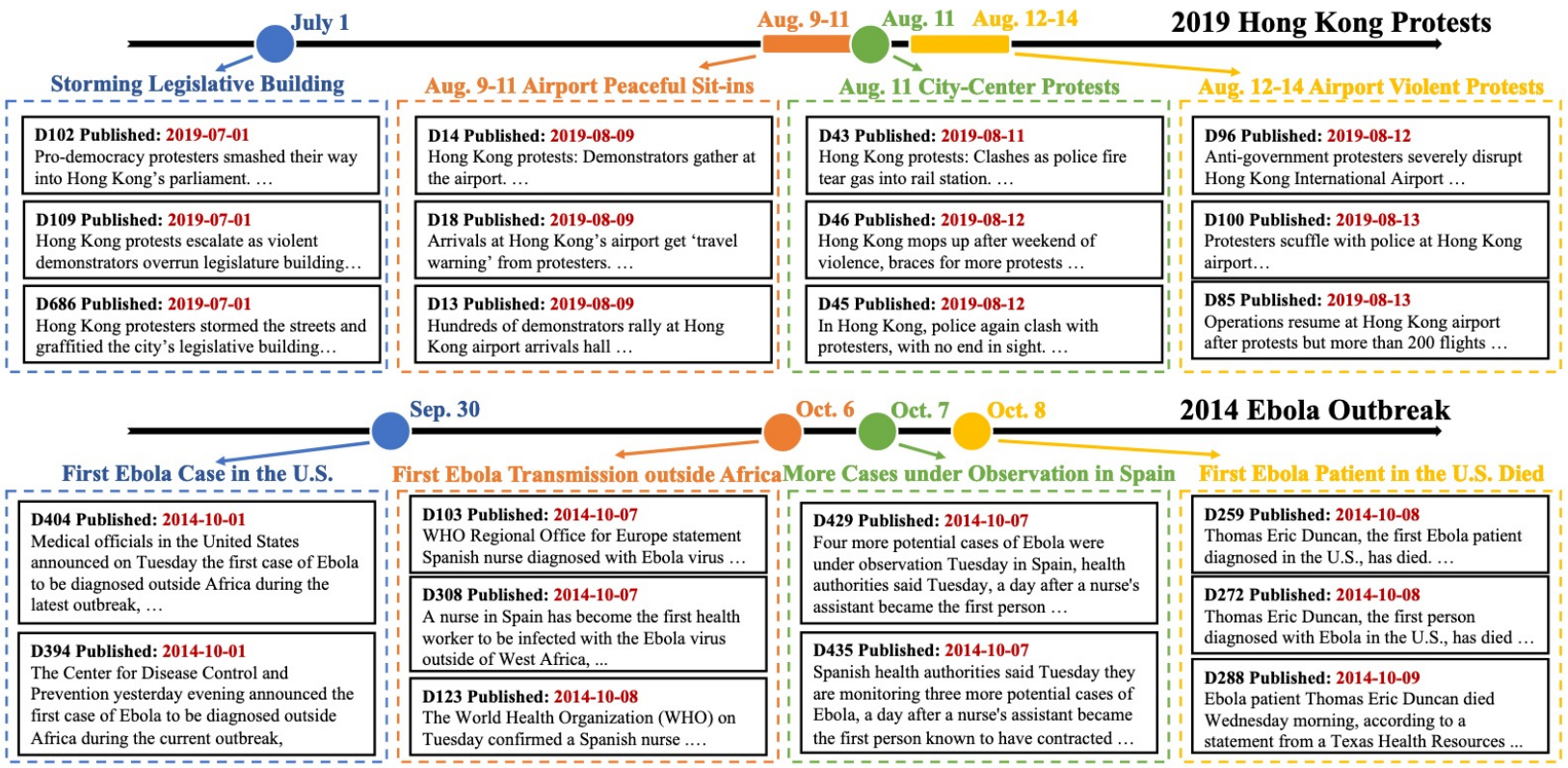


Figure source: Unsupervised Key Event Detection from Massive Text Corpora, Yunyi Zhang, Fang Guo, Jiaming Shen, Jiawei Han, KDD 2022

Events

Reference: https://paperswithcode.com/task/event-extraction
https://paperswithcode.com/task/event-extraction

Event Extraction

80 papers with code • 7 benchmarks • 10 datasets

Determine the extent of the events in a text.

Other names: Event Tagging; Event Identification

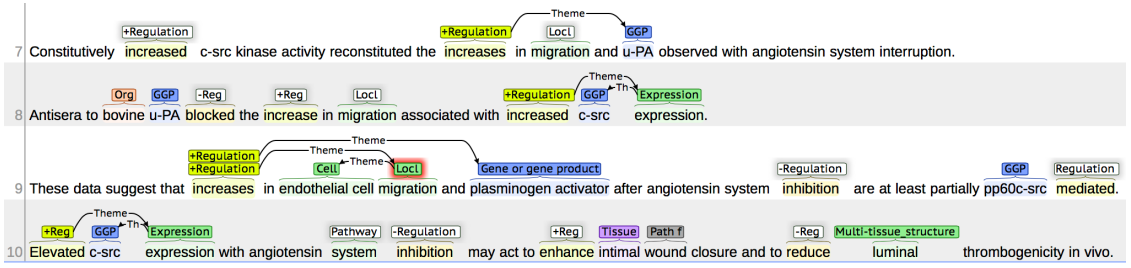
Benchmarks

Add a Result

These leaderboards are used to track progress in Event Extraction

Trend	Dataset	Best Model	Paper	Code	Compare
	GENIA	DeepEventMine			See all
	GENIA 2013	DeepEventMine			See all
	Infectious Diseases 2011 (ID)	DeepEventMine			See all
	Cancer Genetics 2013 (CG)	DeepEventMine			See all
	Epigenetics and Post-translational Modifications 2011 (EPI)	DeepEventMine			See all

Reference: https://github.com/aistairc/DeepEventMine



As on 19 Oct 2022

Aggregation of Events in a Document

- Time
  - By hour, day
- Location
  - By city, county, state, country
- Entities
  - IBM acquired …
  - Red Hat acquired …

Reading Material

- Unsupervised Key Event Detection from Massive Text Corpora, Yunyi Zhang, Fang Guo, Jiaming Shen, Jiawei Han, KDD 2022
  - Paper: https://arxiv.org/abs/2206.04153
  - Code: https://github.com/yzhan238/EvMine
- LnEx
  - Paper: Hussein S. Al-Olimat, Krishnaprasad Thirunarayan, Valerie Shalin, and Amit Sheth. 2018. Location Name Extraction from Targeted Text Streams using Gazetteer-based Statistical Language Models. In Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018), pages 1986–1997. Association for Computational Linguistics.
  - Code: https://github.com/halolimat/LNEx

Location Extraction

- Statistical Inference via n-gram Models
  - Unigrams: Texas, ave
  - Bigram: "Texas ave"

Input: "texas ave is closed"

- Gazeteer Augmentation and Filtering

Given a location name $t_1 \ldots t_n$, we retain t_1 and t_n while varying $t_2 \ldots t_{n-1}$.

  - Only if: t_n is location category: {Building, School, Hospital, Airport, …}
  - "City College of New York" =>
    - {City College, City College of NY}
  - "Balalok Matriculation Higher Secondary School" => {Balalok School, Balalok Secondary School,…}

Uniqueness:
- Method specialized for Twitter text stream
- Aligns with spatial features in Open Street Map

Source: LnEx paper

Predicting Missing Attributes of Events

US Patent: https://patents.google.com/patent/US10296833B2/en

Motivation and Setting

- Given an event, estimate the value of missing attributes. Running example: number of people who may attend it. Example : Book Fair event, Automobile expo.
- Uses
  - Helps organizers prepare for the event
  - Helps host (venue managers, city managers) prepare for the event
  - Helps attendees maximize their experience from the event

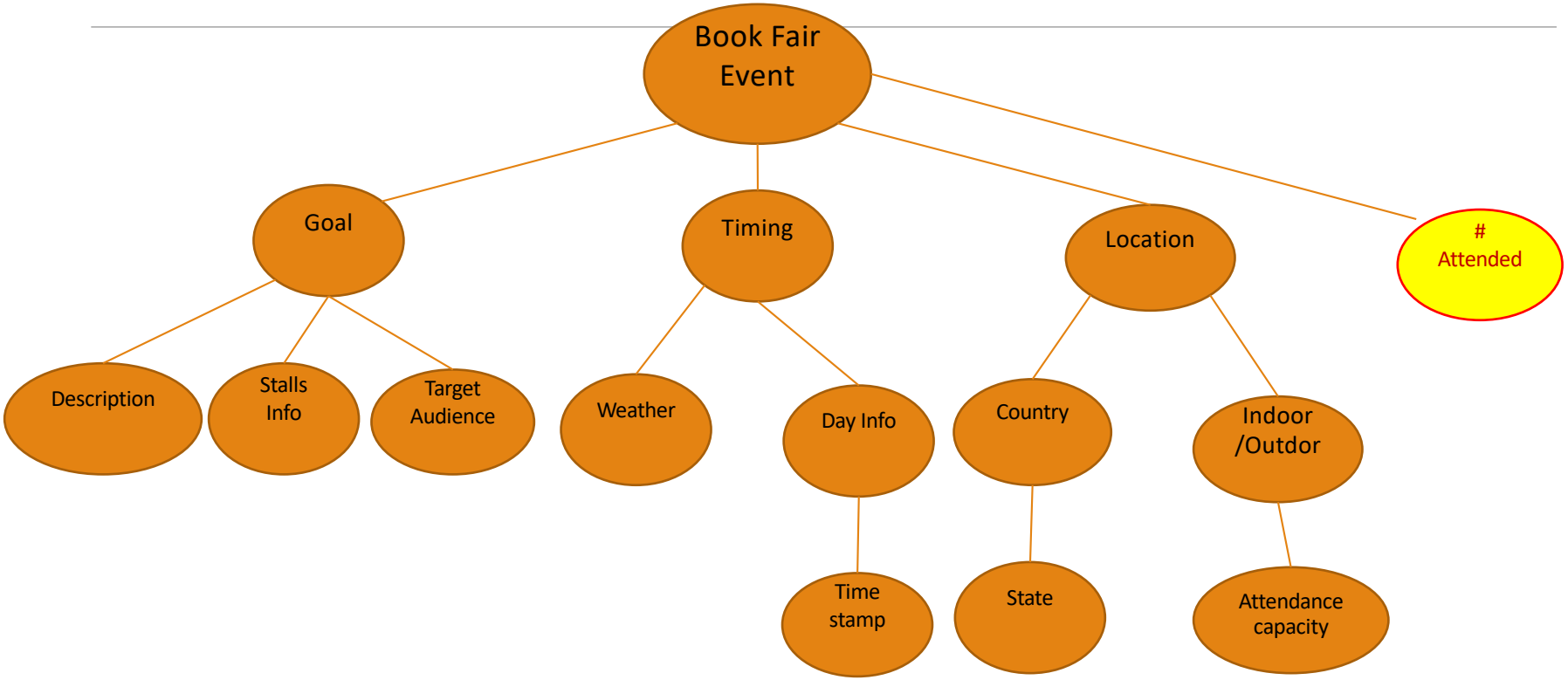
Dataset

- Delhi Book Fair @ Pragati Maidan 2014,2013 ...
- Delhi Automobile Fair @ Pragati Maidan 2014,2013..
- Delhi Automobile Fair @ Ramlila Maidan 2014,2013..
- Madras Book Fair @ YMCA 2014,2013 ...
- Kolkata Automobile Fair @ Milan Mela Ground 2014,2013 ...
- Kolkata Book Fair @ Milan Mela Ground 2014,2013 ...
- London Book Fair @ Olympia 2014,2013 ...
- … 2019

Input : Name - Delhi Book Fair 20xx
Date – March 14-21 20xx,
Goal – The YYrd New Delhi World Book Fair XXX will be held from February aa-bb..
Location - Pragati Maidan, Delhi

Output : # attendees

Book Fair Event Ontology Subset



Data in Example Setting

Delhi Book Fair @ Pragati Maidan 2014,2013 ...
Delhi Automobile Fair @ Pragati Maidan 2014,2013..
Delhi Automobile Fair @ Ramlila Maidan 2014,2013..
Madras Book Fair @ YMCA 2014,2013 ...
Kolkata Automobile Fair @ Milan Mela Ground 2014,2013 ...
Kolkata Book Fair @ Milan Mela Ground 2014,2013 ...
London Book Fair @ Olympia 2014,2013 ...

Event Model Attribute (Concept)	Mandatory/ Opt	Similarity Function	Weights	Comments	Example T*	Example 1	Example 2
Name	M	String Comparison	0.1		Delhi Book Fair 2015	Delhi Book Fair 2014	Kolkatta Book Fair 2014
Description	M	String Comparison	0		…	…	…
Event Type	O	Enum comparison	0.3	Weigh type highly if present	-	-	-
Location	M	Geo- comparison	0.3	Weigh type highly if present	Pragati Maidan, New Delhi	Pragati Maidan, New Delhi	Milan Mela Ground, Kolkatta
Start Date	O	Date comparison	0.05		Feb 2015	Aug 2014	-
End Date	O	Date comparison	0.05		Feb 2015	Aug 2014	-
Attendance	O	Math Subtraction	0.2	Weigh type highly if present	?	50000	70000
Comments	O	String Comparison	0		-	-	-

High-level Solution Steps

- Define ontology for event type
- Find similarity between input event events in dataset using ontology features
  - Use weighing of features
- Rank events based on similarity and filter
- Use regression methods to estimate for missing value

Details: https://patents.google.com/patent/US10296833B2/en

Estimation of Attendance at NFL

- Dataset: https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-02-04/readme.md
- Attendance estimation:
  - Repository: https://github.com/karan109/superBowl/
  - Estimation: https://github.com/karan109/superBowl/blob/master/code/test.ipynb
    - 4.5% mean error
    - Variance based on team and match timings

More Predictions About Recurrent Events (RE)

- RE Examples: conferences, film festivals, sports championships
- Predict
  - Occurrence dates of recurrent events
  - Schedules of these events
- Performance
  - System beats humans in predicting future occurrences of recur-rent events by significant margins

Problem

Inputs:

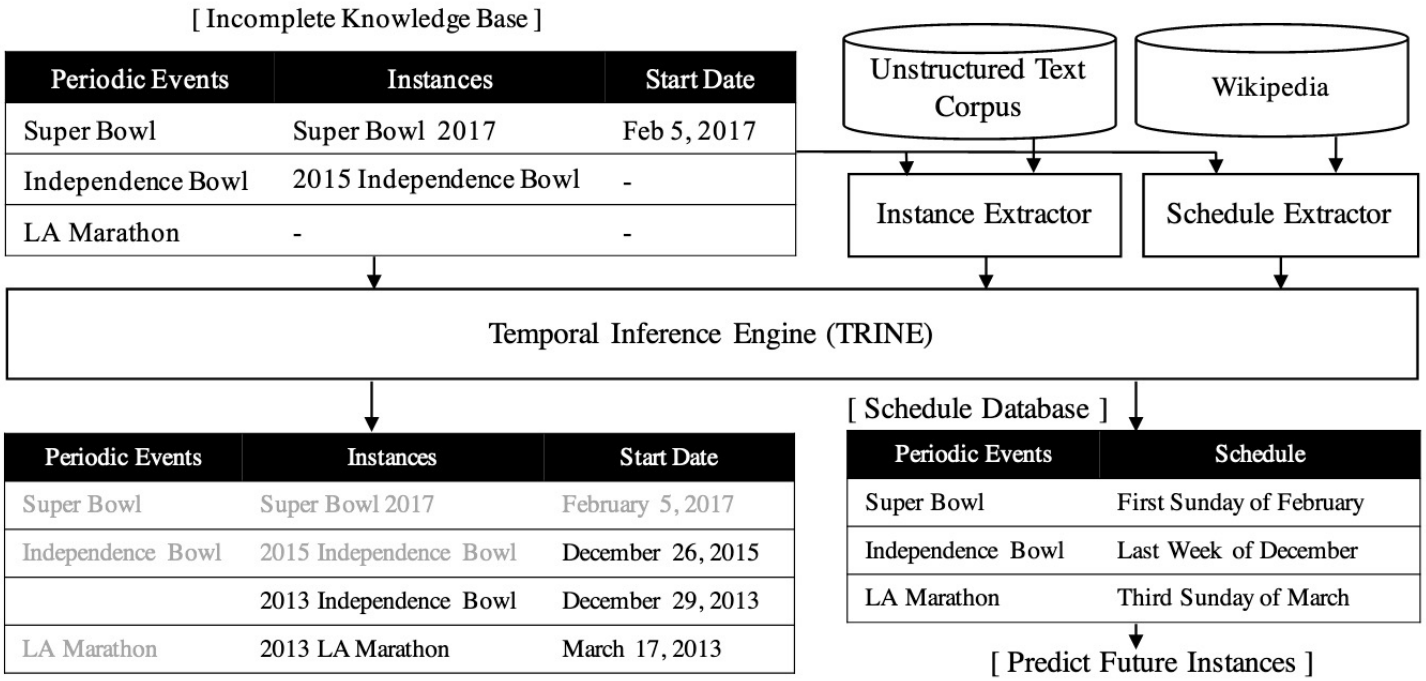
- Recurrent events with a set of instances
- Structured knowledgebase (K) of recurrent events – instances and dates, possibly incomplete. (e.g., database)
- Corpus (C) on untrusted text about recurrent events (e.g., Wikipedia)

Outputs:

- infer schedules for each recurrent event in K
- populate missing information in K
- predict the occurrence date for future instances

Inferring Temporal Knowledge for Near-Periodic Recurrent Events, Dinesh Raghu, Surag Nair, Mausam. International Joint Conference on Artificial Intelligence, (IJCAI). Stockholm, Sweden. July 2018, https://www.ijcai.org/Proceedings/2018/0598.pdf ; **Code**: https://github.com/dair-iitd/trine

TRINE: Predictions About Events



Inferring Temporal Knowledge for Near-Periodic Recurrent Events, Dinesh Raghu, Surag Nair, Mausam. International Joint Conference on Artificial Intelligence, (IJCAI). Stockholm, Sweden. July 2018, https://www.ijcai.org/Proceedings/2018/0598.pdf ; **Code**: https://github.com/dair-iitd/trine

Solution Building Blocks and Results

Main Blocks

- Grammar for recurrent events
- Schedule extractor (For K, extract from C)
- Instance extractor (from C)

Month Season (MS)	Date (DT)	Day/Week (DW)	Month/Season Modifier (MS_{mod})	Day/Week Modifier (DW_{mod})
JAN	1	MON	EARLY	1st
...	2	TUE	MID	2nd
DEC	3	...	LATE	3rd
SUMMER	4	SAT		4th
...	...	SUN		LAST
WINTER	31	WEEK		

Table 1: Atomic Elements for Schedule Representation

$$S \rightarrow C \mid (C \text{ or } S)$$

$C \rightarrow MS$ $\quad C \rightarrow MS\ DT$

$C \rightarrow MS_{mod}\ MS$ $\quad C \rightarrow DW_{mod}\ DW \text{ of } MS$

Figure 2: CFG Rules for Schedule Space Representation

	MRR	Acc@1	Acc@3	Acc@5
Mode Baseline	0.096	0.019	0.094	0.264
Mean Baseline	0.147	0.075	0.17	0.32
AMT	0.308	0.241	0.366	0.41
TRINE	**0.388**	**0.292**	**0.434**	**0.575**

Table 2: Future occurrence date prediction performance for 2016 instances of 106 recurrent events

Inferring Temporal Knowledge for Near-Periodic Recurrent Events, Dinesh Raghu, Surag Nair, Mausam. International Joint Conference on Artificial Intelligence, (IJCAI). Stockholm, Sweden. July 2018, https://www.ijcai.org/Proceedings/2018/0598.pdf ; **Code**: https://github.com/dair-iitd/trine

Additional Assessment of Events

Medical:

- **29K** COVID-19 new cases were reported on **April 30, 2020**
- **1.3M** cases of COVID-19 were detected in the **US**
- CDC, Washington State Report First COVID-19 Death [Feb 29, 2020]
- **29K** COVID-19 new cases were reported in the **US** on **April 30, 2020**
- 'There will be a day of reckoning' for those trying to price-gouge doctors and nurses: New Jersey governor **[1 Apr 2020]**

Financial:

- Xerox drops $34 bn deal to buy bigger rival HP **[1 Apr 2020]**
- Xerox walks away from $35 billion hostile bid for HP **[1 Apr 2020]**
- Tricia Fitzmaurice Takes Nat'l Security Program Director Role at Red Hat **[1 Apr 2020]**

About the frequency and spread of events

About the magnitude of an event's impact:

- Quantitative numbers
- Sentiments
- Superlatives

Relative intensity for similar events can be calibrated

- Numeric size: $100M v/s $34B
- Discounting by time and distance

Sound Event Detection (SED)

- Event in sound (audio) media with temporal start and end time
  - Recognizing overlapping sound events called polyphonic SED

Benchmarks

Add a Result

These leaderboards are used to track progress in Sound Event Detection

Trend	Dataset	Best Model	Paper	Code	Compare
	DESED	FDY-CRNN			See all
	L3DAS21	PHC SEDnet n=2			See all
	Mivia Audio Events	DENet			See all
	Mivia Road Events	DENet			See all

Reference: https://paperswithcode.com/task/sound-event-detection

As on 19 Oct 2022

Lecture 18: Concluding Comments

- We looked at Event detection
  - Spatio-temporal analysis
  - Many practical applications
- Related sound event detection

Concluding Segment

About Next Lecture – Lecture 19

Lecture 19 Outline

- Topic analysis
  - Latent Dirichlet Allocation (LDA)
  - Latent Semantic Analysis (LSA)
- Usage