ISHER:
Integrated Social History Environment for Research

Sophia Ananiadou

National Centre for Text Mining

Aims

- Search system
  - End-users: social history researchers
  - News reports on social unrest
    - New York Times; 1.8 million articles (1987-2007)
  - ACE2005 corpus
    - NE and event annotations on news articles
- Develop and evaluate analytics
  - Access to corpora and searchable data sources
  - Analysis engines for entities and events
  - Curate annotations using UIMA-based platforms (Argo, U-Compare)
- Extract meta-knowledge for social history events

Infrastructure

- An annotation text mining workflow for recognising events in the social domain in UIMA / U-Compare / Argo

- An environment for searching documents annotated by the workflow above.

ISHER: value-added services

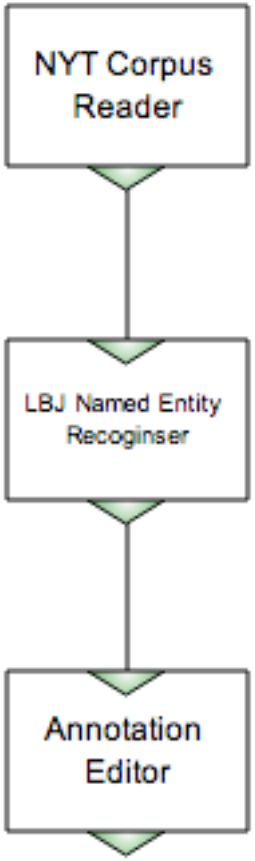
- Search service
  - Faceted search over entity and event annotations
  - Clustering of similar news items
  - Cluster visualisation
- Social unrest database curation pipeline
  - Filter relevant documents
  - Apply entity and event analytics
  - Edit annotations
  - Add to searchable index

DEMO

http://www.nactem.ac.uk/ISHER/

Curation pipelines in Argo

- Start with a collection reader.
  - Corpus reader for internal use
- One or more analytic engines
- Annotation editor
  - to correct/augment
- A ‘consumer’ to collect annotations for indexing

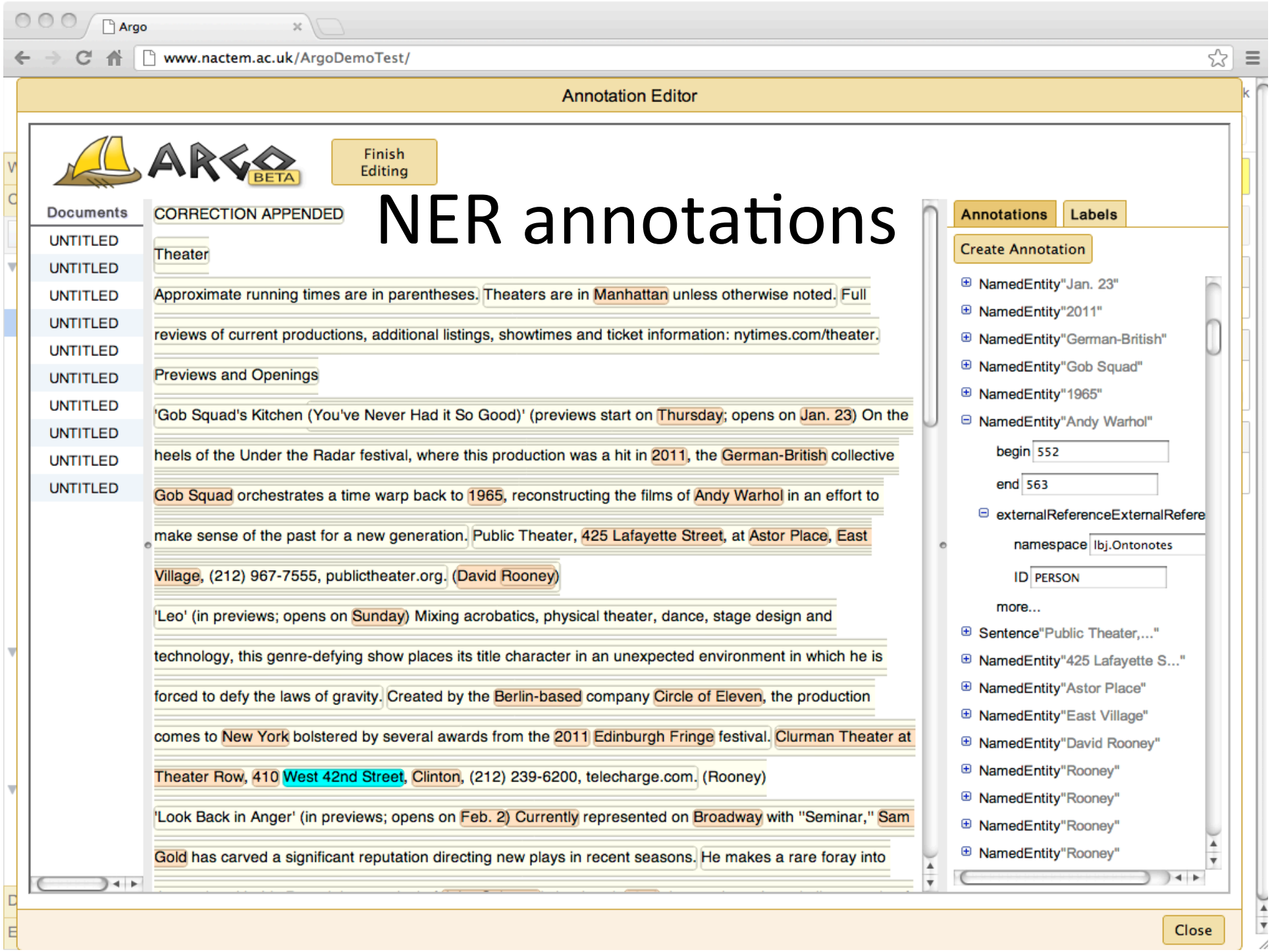


http://www.nactem.ac.uk/argo/

Named Entity Recognition

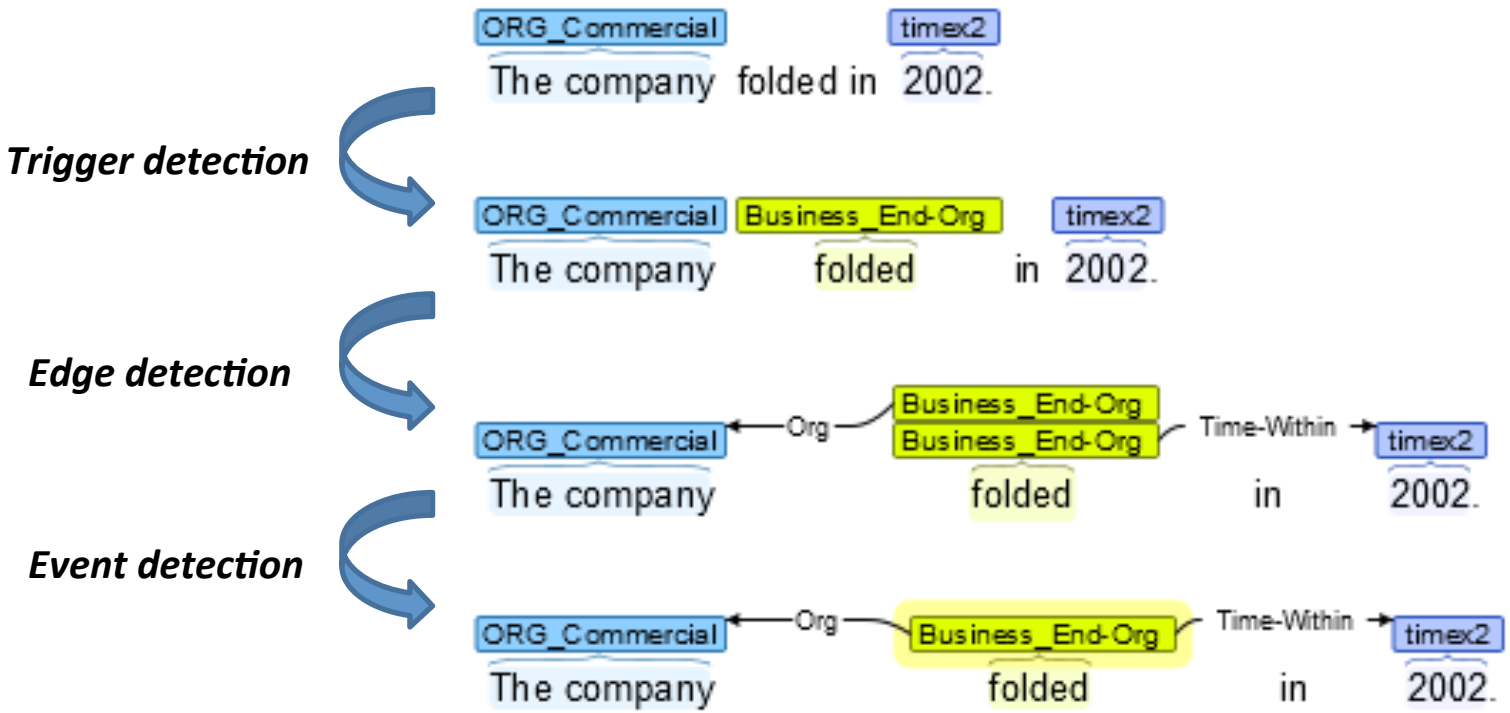
- Leveraging NER analytics from University of Illinois
  - http://cogcomp.cs.illinois.edu/page/tools
- Wrapped for Argo (UIMA)
- Configuration to output wide variety of NE and related expressions, e.g.:

  TIME LAW LANGUAGE PERCENT PRODUCT ORDINAL LOC PERSON WORK_OF_ART DATE QUANTITY ORG CARDINAL


Argo
www.nactem.ac.uk/ArgoDemoTest/
Annotation Editor
ARGO BETA
Finish Editing
NER annotations
Documents
UNTITLED
UNTITLED
UNTITLED
UNTITLED
UNTITLED
UNTITLED
UNTITLED
UNTITLED
UNTITLED
UNTITLED
CORRECTION APPENDED
Theater
Approximate running times are in parentheses. Theaters are in Manhattan unless otherwise noted. Full reviews of current productions, additional listings, showtimes and ticket information: nytimes.com/theater.
Previews and Openings
'Gob Squad's Kitchen (You've Never Had it So Good)' (previews start on Thursday; opens on Jan. 23) On the heels of the Under the Radar festival, where this production was a hit in 2011, the German-British collective Gob Squad orchestrates a time warp back to 1965, reconstructing the films of Andy Warhol in an effort to make sense of the past for a new generation. Public Theater, 425 Lafayette Street, at Astor Place, East Village, (212) 967-7555, publictheater.org. (David Rooney)
'Leo' (in previews; opens on Sunday) Mixing acrobatics, physical theater, dance, stage design and technology, this genre-defying show places its title character in an unexpected environment in which he is forced to defy the laws of gravity. Created by the Berlin-based company Circle of Eleven, the production comes to New York bolstered by several awards from the 2011 Edinburgh Fringe festival. Clurman Theater at Theater Row, 410 West 42nd Street, Clinton, (212) 239-6200, telecharge.com. (Rooney)
'Look Back in Anger' (in previews; opens on Feb. 2) Currently represented on Broadway with "Seminar," Sam Gold has carved a significant reputation directing new plays in recent seasons. He makes a rare foray into
Annotations
Labels
Create Annotation
NamedEntity"Jan. 23"
NamedEntity"2011"
NamedEntity"German-British"
NamedEntity"Gob Squad"
NamedEntity"1965"
NamedEntity"Andy Warhol"
begin 552
end 563
externalReferenceExternalRefere
namespace lbj.Ontonotes
ID PERSON
more...
Sentence"Public Theater,..."
NamedEntity"425 Lafayette S..."
NamedEntity"Astor Place"
NamedEntity"East Village"
NamedEntity"David Rooney"
NamedEntity"Rooney"
NamedEntity"Rooney"
NamedEntity"Rooney"
NamedEntity"Rooney"
Close

Event Extraction Text Mining Workflow

- Pipeline system with machine-learning-based modules



http://www.nactem.ac.uk/EventMine/

Data set

- ACE 2005 Multilingual Training Corpus
  - 599 English articles
    - 5,349 events: 8 types, 33 subtypes (e.g., *Life-Marry*)
    - 9,793 roles: 35 role types (e.g., *Agent*)
    - 61,321 entities/values (~ event arguments)
      - 54,824 entities: 7 types, 45subtypes (e.g., *ORG-Media*)
      - 5,469 timex2 elements: 1 type
      - 1,028 values: 5 types, 5 subtypes (e.g., Numeric-Money)

Event Extraction Performance

System	Trigger (F-Score)	Role (F-Score)
EventMine	**71.0**	**52.1**
EventMine Service	64.5	46.9
Liao et al., 2010	68.8	44.6

- EventMine produces comparable results with the state-of-the-art system (note: different test set)
- EventMine Service does **not** use
  - Entity surface (e.g., "Bush" for PER) for generalisation
  - Entity subtypes/attributes which are rarely automatically recognised

Enriching the ACE corpus with meta-knowledge

Interpretation of events

- The interpretation of events varies according to context and textual features. A event may:
  - represent a fact, a specific event, an opinion/analysis, a hypothetical situation, recommendation, etc.
  - be presented as the author's own knowledge/point of view, or that of a third party
  - be expressed together with a level of certainty (e.g., speculation)
  - Represent a situation in the past, an ongoing situation, or something that will happen in the future
  - be negated, i.e., there is an indication that the event did not happen
- *meta-knowledge*

Meta-knowledge examples (1)

- A fact
  - *In all,* ***two million Americans have lost their jobs under President Bush so far****.*
- A specific past event
  - *Even as the secretary of homeland security was putting his people on high alert last month,* ***a 30-foot Cuban patrol boat with four heavily armed men landed on American shores****.*

Meta-knowledge examples (2)

- A hypothetical event
  - *It could swell to as much as $500 billion if **we go to war in Iraq**.*
- An opinion of a third party
  - Dan Snyder of Baden, Pennsylvania writes, "***Bush should torture the al Qaeda chief operations officer***."
- A speculative (low certainty) analysis made by a third party
  - ***John Paul II might retire at the end of this year,*** *a Belgian cardinal says*

Enriching event annotation

- Enriching event annotation with meta-knowledge information allows more sophisticated event extraction systems to be trained
- Additional search criteria can be specified e.g.
  - Find only events that represent known facts
  - Find only events which are reported with high or complete certainty

Meta-knowledge annotation (1)

- Different types of meta-knowledge can be encoded as different attributes or dimensions annotated for each event
- ACE includes some event attributes relating to meta-knowledge
  - Tense
  - Polarity
  - Modality – real or abstract event
  - Specificity – a single specific event or a generic event

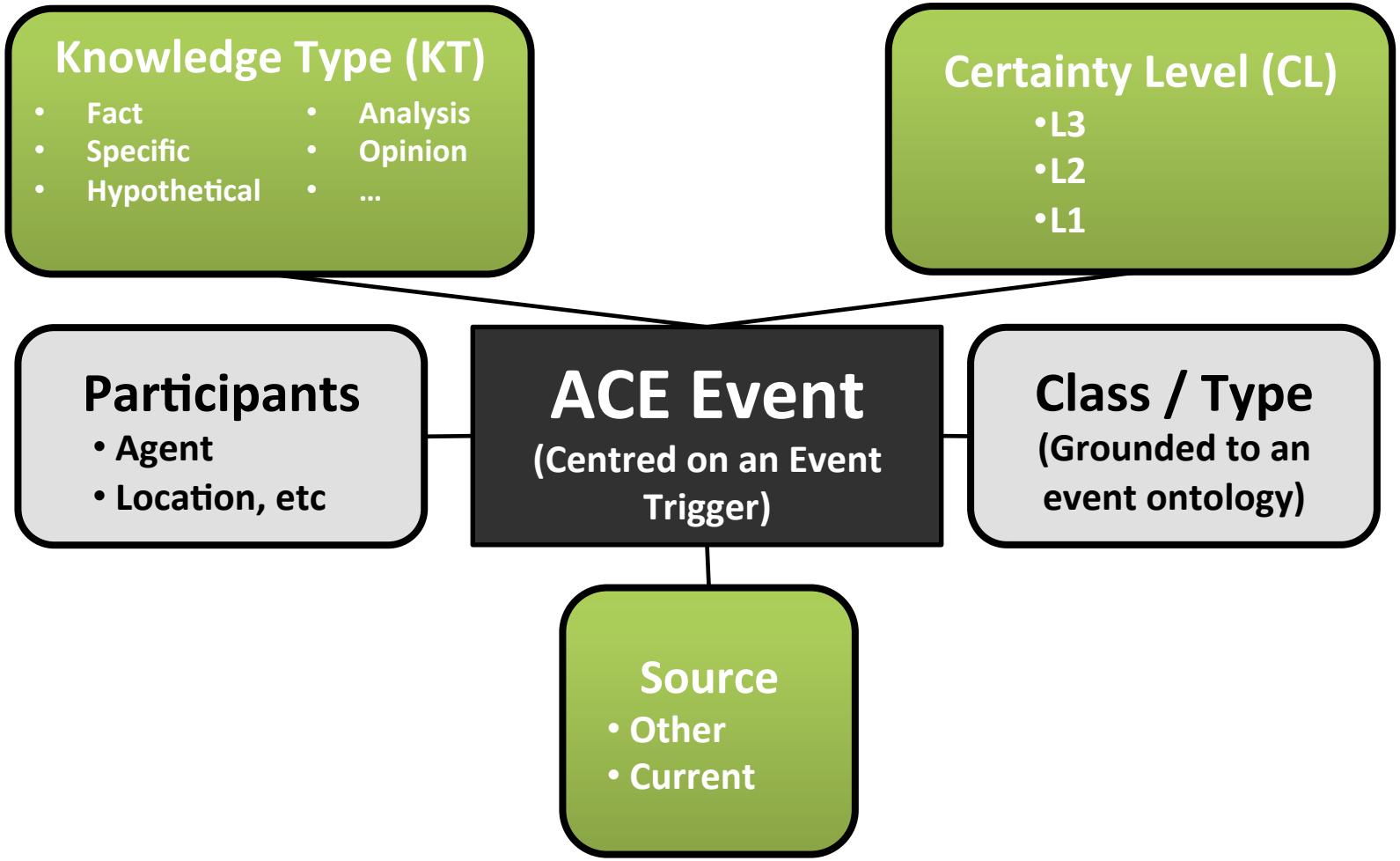
Meta-knowledge annotation (2)

- The existing ACE scheme misses certain information about events
  - e.g., certainty level, point of view (author or third party)
- It is unable to make certain distinctions about abstract events
  - Analyses, opinions, hypothetical situations

NaCTeM Meta-knowledge scheme

- Originally designed to be applied to biomedical text
- Allows fine-grained distinctions to be made between event interpretations
- Can be applied consistently by different annotators
  - Average 0.88 Kappa
- Systems have been trained to predict event meta-knowledge automatically
  - Miwa, M., Thompson, P., McNaught, J., Kell, D. B. and Ananiadou, S. (2012). Extracting semantically enriched events from biomedical literature. *BMC Bioinformatics* 13: 108
- Adapt the scheme to and apply it the ACE 2005 corpus

Proposed Scheme



Remaining work

- Expand from ACE data to NYT
- Enrich with meta-knowledge of events
- Customise towards social unrest
- Finalise search system

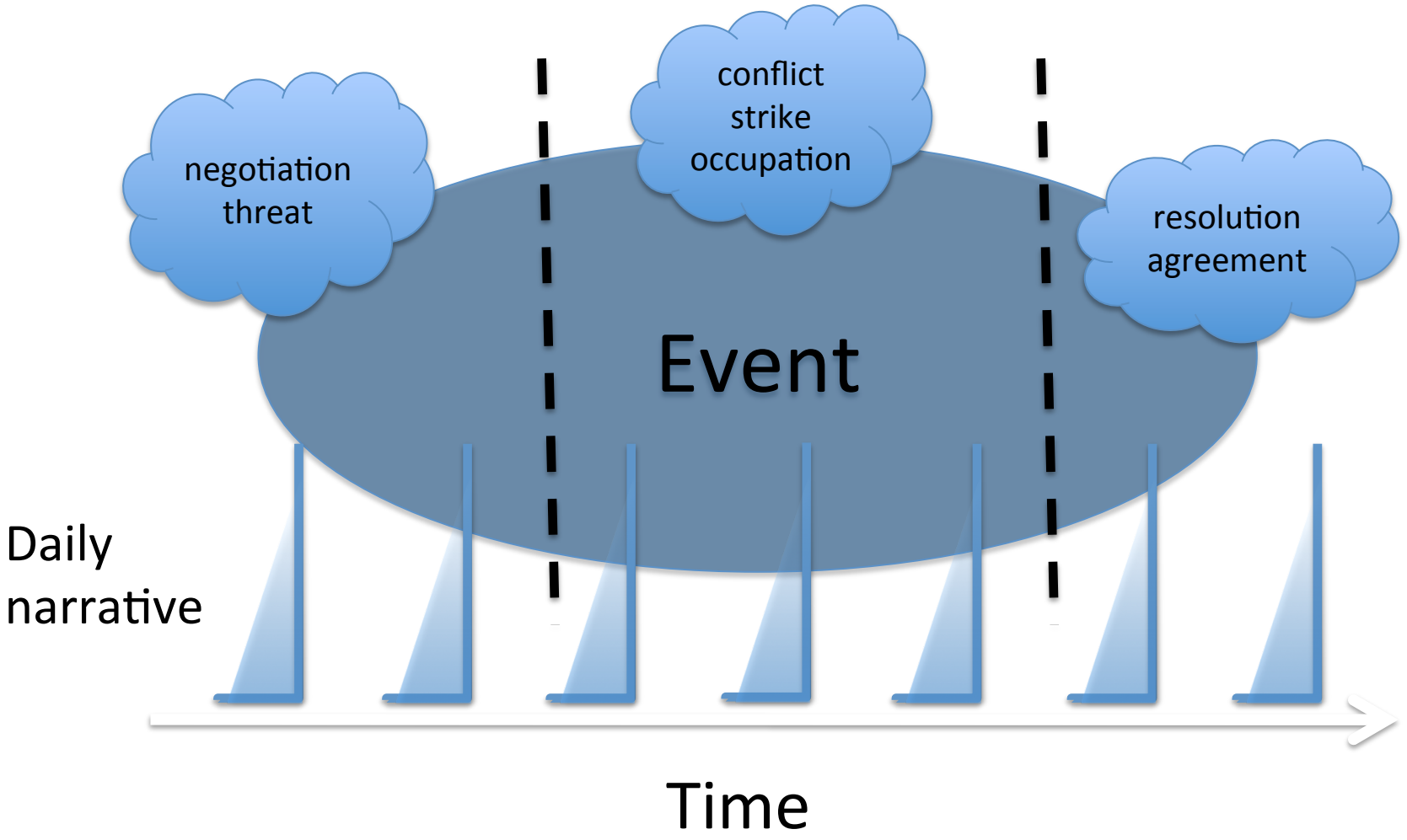
The University of Manchester

Beyond reported history: Strikes that never happened

Antal van den Bosch, Kalliopi Zervanou,
Martha van den Hoven

Digging Into Data: ISHER
Centre for Language Studies, Radboud University Nijmegen

Events in time



Bedrijf	Leidsche Katoenmaatschappij
Plaatsen	Leiden (Zuid-Holland)
Beroep	katoenwever
Eis	tegen loonsverlaging
Sector	Industrie/bouw
Soort actie	Staking
Type staking	Klassiek
Karakter	Vakbond
Resultaat	Verlies
Datum	22 februari 1922
Duur	133 dagen
Verslag	De bonden wilden eind mei een eind aan de strijd, maar de stakers weigerden dit. Waarnemend burgemeester Van der Lip deed een poging tot bemiddeling. Half augustus was vrijwel iedereen weer aan het werk. Veel sympathie van de bevolking. Van de stakers waren er 213 bondslid. In de vergadering van 27 juni van de Arbeidsraad bedankte de directie de bazen voor hun hulp tijdens de staking.
Aantallen	Stakers (hoogste aantal stakers): 265 Gestaakte dagen (totaal niet-gewerkte mensdagen): 30210 Onvrijwillige stakers (werknemers die door de actie niet verder kunnen werken): 0 Verloren dagen onvrijwillige stakers (arbeidsdagen dat de niet stakers niet konden werken): 0 Aantal betrokken bedrijven (aantal betrokken bedrijven): 1

Newspaper archives

- Koninklijke Bibliotheek in The Hague
  - Newspapers from 1618 to 1995
  - Over 1 million pages online, target 8 million
  - Searchable on title, date, words
- Cf. *Forces of labor*, Silver (2003)

Searching for articles on strikes

- Create query (model / classification rule)
- Most strikes in 1910-1940
- Search terms: staking, staken
  - Earlier: *strike, grève, bollejeije, laveij*
  - Later: *werkonderbreking, stiptheidsactie*
- From strike database record:
  - date and length
  - names of companies, unions, locations, occupations, sectors

Bedrijf	Leidsche Katoenmaatschappij
Plaatsen	Leiden (Zuid-Holland)
Beroep	katoenwever
Eis	tegen loonsverlaging
Sector	Industrie/bouw
Soort actie	Staking
Type staking	Klassiek
Karakter	Vakbond
Resultaat	Verlies
Datum	22 februari 1922
Duur	133 dagen
Verslag	De bonden wilden eind mei een eind aan de strijd, maar de stakers weigerden dit. Waarnemend burgemeester Van der Lip deed een poging tot bemiddeling. Half augustus was vrijwel iedereen weer aan het werk. Veel sympathie van de bevolking. Van de stakers waren er 213 bondslid. In de vergadering van 27 juni van de Arbeidsraad bedankte de directie de bazen voor hun hulp tijdens de staking.
Aantallen	Stakers (hoogste aantal stakers): 265
	Gestaakte dagen (totaal niet-gewerkte mensdagen): 30210
	Onvrijwillige stakers (werknemers die door de actie niet verder kunnen werken): 0
	Verloren dagen onvrijwillige stakers (arbeidsdagen dat de niet stakers niet konden werken): 0
	Aantal betrokken bedrijven (aantal betrokken bedrijven): 1
	Aantal acties (aantal acties): 1

Example query

```
SELECT all articles
CONTAINING THE WORDS
stak?n* AND
(blokband OR Amsterdam OR
  taxichauffeur)
AND artikeldatum BETWEEN 6 apr 1937 – 7
                         AND 9 apr 1937 + 3
```

Results

De staking der Amsterdamsche taxichauffeurs

Mededeelingen van werkgeverszijde

DE RIJKSBEMIDDELAAR VRAAGT INLICHTINGEN.

Naar wij vernemen, heeft de rijksbemiddelaar, mr dr S. de Vries Czn, de werkgevers in het taxibedrijf te Amsterdam verzocht hedenmiddag een afgevaardigde naar Den Haag te zenden om hem in te lichten omtrent de staking in het blokbandtaxi-bedrijf te Amsterdam.

De voorzitter van het bestuur der taxi-centr de heer C. J. van Leusden, heeft in verba met de staking medegedeeld, dat tot Woensd avond noch het bestuur der Taxi-centrale, n de daarbij aangesloten ondernemers, van chauffeurs eenig verzoek hebben ontvangen in verband met de verhooging der tarieven loonen te herzien. De staking is Dinsdagmid uitgebroken zonder dat te voren overleg is pleegd en ondanks het feit, dat er tusschen werkgevers in het blokband-taxibedrijf en vier groote organisaties van transportarbeid waarbij de chauffeurs zijn aangesloten, een d het gemeentebestuur goedgekeurde collecti arbeidsovereenkomst bestaat, waarvan art bepaalt, dat beide partijen zich verbinden durende den duur der overeenkomst g werkstakingen of uitsluitingen tegen elkan te proclameren of uit te voeren of te doen voeren.

Deze staking moet dus, volgens het oord der werkgevers, als een wilde staking wor beschouwd.

De staking der Amsterdamsche taxichauffeurs

Rijksbemiddelaar verklaart zich onbevoegd verder van het geschil kennis te nemen

DE GELDENDE ARBEIDSOVEREENKOMST VAN KRACHT.

De Rijksbemiddelaar in het 2e district, mr S. de Vries Czn, heeft gisteren in het Departement van Sociale Zaken een conferentie gehad met de werkgevers- en werknemersorga trokken bij het conflict in het sta taxi-bedrijf te Amsterdam.

Allereerst is daarbij komen vast de collectieve arbeids-overeenkoms op 1 Augustus 1936, thans nog geld

Artikel 3 dier C.A.O. bepaalt, dat tuerende partijen en de leden der organisaties zich verbinden, gedurer van de overeenkomst geen werksta sluiting tegen elkander te proclam te voeren of te doen uitvoeren.

De bedoeling der staking is, om duur der C.A.O. daarin wijzinge brengen, op grond dat gevreesd w tariefsverhooging een nadeeligen hebben op de loonen.

Naar de meening van den Rijks is door de stakers niet den juiste wandeld. Immers in de C.A.O. is g gelegd tusschen den ritprijs en he

DE AMSTERDAMSCHE TAXI-STAKING GEëINDIGD.

Tariefsverhooging blijft gehard haafd.

Concessies aan de chauffeurs.

De stakende Amsterdamsche taxi-chauffeurs hebben op een gistermiddag in „Krasnapolsky" gehouden vergadering besloten de staking op te heffen, op basis van het bemiddelingsvoorstel, dat een concessie aan de werknemers genoemd kan worden. Des avonds heeft de avondploeg zich bij de garages gemeld en de chauffeurs reden met hun wagens naar de standplaatsen. Amsterdam was weer uit den taxi-nood!

De getroffen regeling komt in het kort

dhaafd
ats van
wordt
en voor
onder
menko-
t stand
oor het

„Sit Down" als Reclame!

In een slagerij in de Kinkerstraat te Amsterdam is door het personeel, vermoedelijk met medewerking van den patroon, een sit-down-staking in de winkel geënscèneerd. Deze „staking" was, naar A.N.I.P.-Aneta uit Amsterdam seint, klaarblijkelijk bedoeld als reclame!

De politie moest charges uitvoeren om de straat voor het verkeer vrij te houden.

De „staking" werd inmiddels opgeheven.

Discussion

- Straightforward, limited, semi-manual test produces potentially usable results
- Up to historians to value the outcome and the procedure
- To what other research questions does this generalize?
- Underlying model: the narrative of an event over time
- Computational counterfactual history