Statistical information extraction : a tutorial and real-life case study

Marc Vilain
The MITRE Corporation

Presented at ASIAS symposium
July 27-28, 2009

MITRE

What is information extraction?

- **Identification of entities, relationships, and events in language sources**
  - People, places, organizations, vehicles, …
  - Employment, family relations, domicile, …
  - Merger/acquisition, kidnapping, disease outbreak, …
- **Basis for indexing, mapping, database creation, statistical analysis, knowledge discovery …**
- **Application areas**
  - Law enforcement and national security
  - Business intelligence (growth area)
  - Epidemiology mapping

CIIS
Center for Integrated Intelligence Systems

CENTER FOR ENTERPRISE MODERNIZATION

MITRE

ASAP reports: entities

airport

runway

In **MKG** we were given clearance of 'taxi to **runway 14** via **36**.' …
While taxiing to **14** we were told to hold short for landing …
We were confused as to which line was the **hold short line for runway 14** as there were **two hold short lines** in close proximity. …
We unintentionally crossed the **hold short line for runway 14** thinking it was the **hold short for runways 18/36** …
The event occurred due to confusion of the **runway markings**, airport map page, and information [from ATC] …
The **markings for the hold short** and **markings for runway 18/36** did not look right …
I was also preoccupied with running checklists and informing the flight attendant of our immediate departure …

ground markings

CIIS
Center for Integrated Intelligence Systems

CENTER FOR ENTERPRISE MODERNIZATION

MITRE

ASAP reports: events

airplane motion

In MKG we were given clearance of '**taxi to runway 14 via 36**.' …
While **taxiing to 14** we were told to **hold short for landing** …
We were confused as to which line was the hold short line for runway 14 as there were two hold short lines in close proximity. …

incident

We unintentionally **crossed the hold short line for runway 14** thinking it was the hold short for runways 18/36 …

procedure

The event occurred due to confusion of the runway markings, airport map page, and information [from ATC] …
The markings for the hold short and markings for runway 18/36 did not look right …
I was also preoccupied with **running checklists** and **informing the flight attendants** of our immediate departure …

CIIS
Center for Integrated Intelligence Systems

CENTER FOR ENTERPRISE MODERNIZATION

MITRE

ASAP reports: sentence classification

confusion

In MKG we were given clearance of '**taxi to runway 14 via 36**.' …
While **taxiing to 14** we were told to **hold short for landing** …

We were confused as to which line was the **hold short line for runway 14** as there were **two hold short lines** in close proximity. …

responsibility

We unintentionally **crossed the hold short line for runway 14** thinking it was the **hold short for runways 18/36** …

The event occurred due to confusion of the **runway markings**, airport map page, and information [from ATC] …

The **markings for the hold short** and **markings for runway 18/36** did not look right …

I was also preoccupied with **running checklists** and **informing the flight attendants** of our immediate departure …

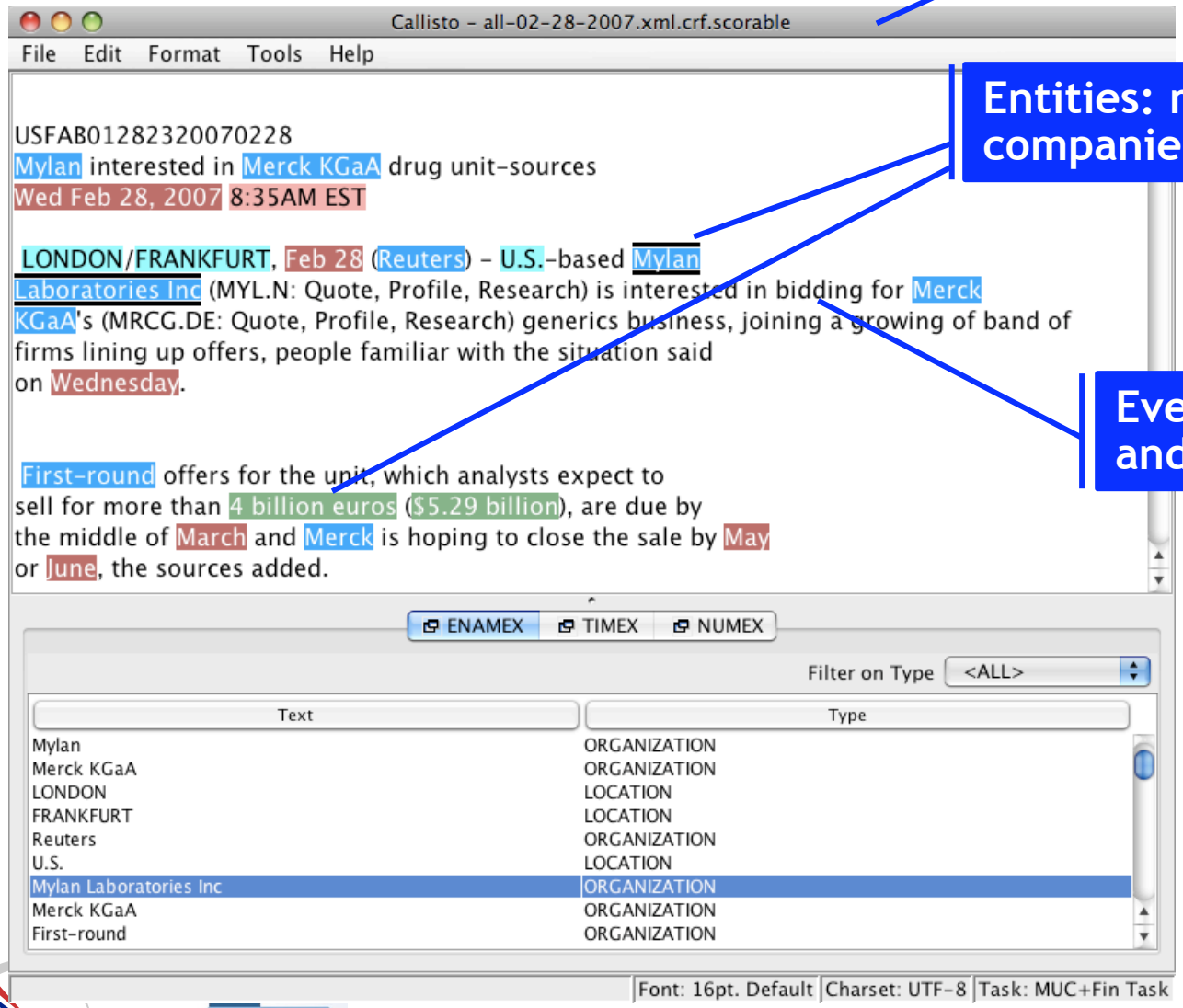
attention

CIIS Center for Integrated Intelligence Systems

CENTER FOR ENTERPRISE MODERNIZATION

MITRE

Case study: Business news



Callisto markup tool

Entities: money, companies, …

Events: mergers and acquisitions

CIIS Center for Integrated Intelligence Systems

CENTER FOR ENTERPRISE MODERNIZATION

MITRE

Approaches to extraction

- **Rule-based**
  - Entities extracted by pattern-matching rules
  - MR. XXX → MR. <PERSON>XXX</PERSON>
  - Pattern rules must be built and ordered by hand
  - To port to new task, rule base must be manually modified; for rule-based tools, this usually requires hiring the vendor
- **Statistical**
  - Entities extracted by statistical evidence-weighing
  - $P(w_0 \approx \text{person} \mid w_{-1} = \text{“MR.”})$
  - Probabilities are estimated from training data
  - To port to new task, retrain estimates with task-specific training data

Probability that current word is person-typed, given that previous word is “MR.”

CIIS
Center for Integrated Intelligence Systems

CENTER FOR ENTERPRISE MODERNIZATION

MITRE

Statistical extraction: Business news how-to

(0) Train baseline system from legacy newswire data

(1) Manually annotate M&A data

Tag 150 news stories (one day of news) to set guidelines

Tag 3 additional news days for training/eval

(2) Create additional training data

Manually tag related news (hot stocks, general biz, …)

Partially auto-tag more M&A news

(3) Train statistical CARAFE models

Implement various tricks to handle auto-tagged data and capture discourse effects

Evaluate performance

CIIS
Center for Integrated Intelligence Systems

CENTER FOR ENTERPRISE MODERNIZATION

MITRE

Technical tricks

- **Resources**
  - Gazetteers capture states/provinces, known first names
  - Other lists capture days of week, months, etc.
- **Long-distance dependencies**
  - "Thomas White … . White was named global head of …"
  - First instance of "White" predicted from context ("Thomas"), but second instance lacks predictive context
  - Use feature-copying trick:
    $\Phi(w)$: $w'_{-1} \in$ *first-names* for some w'=w in the narrative
  - Also effective for banks (Global Credit Union … Global) and companies (Zelcor Widgets Corp. … Zelcor)

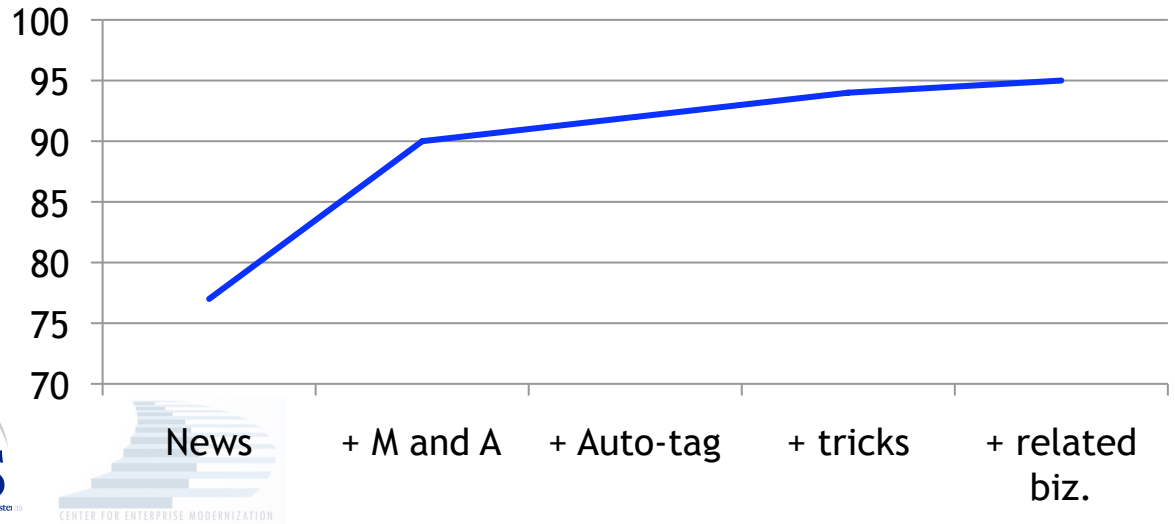
CIIS Center for Integrated Intelligence Systems

CENTER FOR ENTERPRISE MODERNIZATION

MITRE

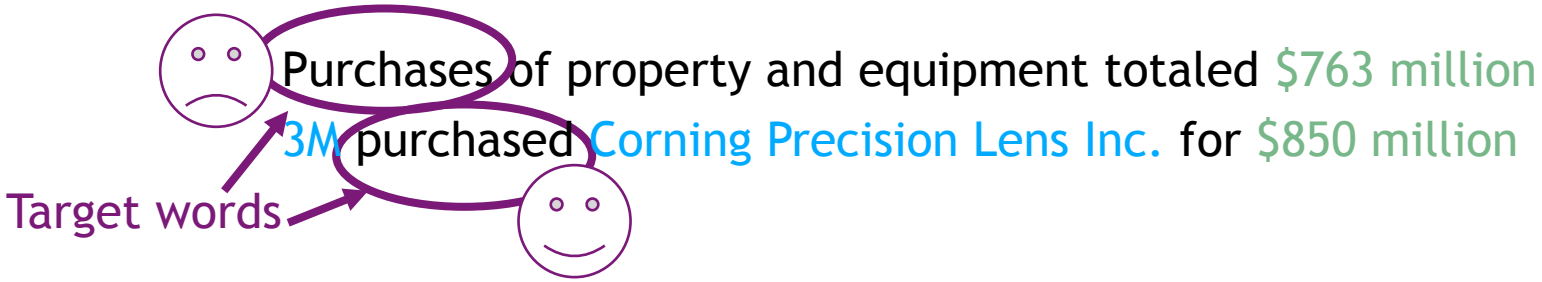
Evaluation

- **Overall performance: F=95 (R=95, P=95)**
- **Performance for specific entities**
  - Organization F=94 (R=94, P=95)
  - Person F=92 (R=92, P=92)
  - Location F=93 (R=92, P=94)
  - Date F=97 (R=98, P=97)
  - Money F=98 (R=97, P=99)

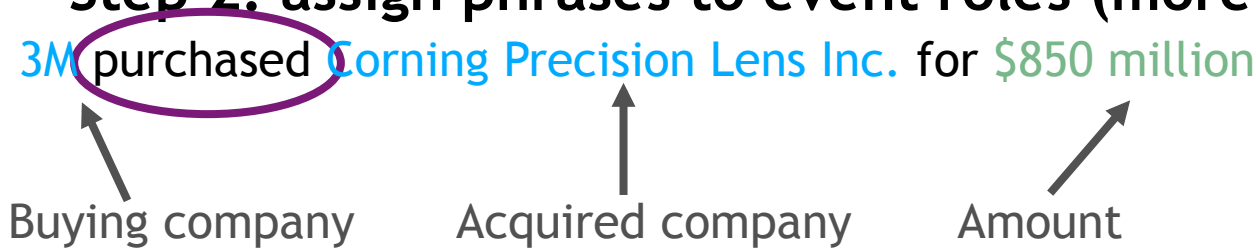


Statistical identification of events

- Multi-stage process based on statistical classifiers
- Step 1: classify target words as to whether they're event-denoting (maximum entropy models)



- Step 2: assign phrases to event roles (more classifiers)



CIIS Center for Integrated Intelligence Systems

CENTER FOR ENTERPRISE MODERNIZATION

MITRE

Conclusions

- **Statistical methods just simply work**
  - Trainable methods allow for rapid adaptation to new tasks
  - Great success with many tasks and many sources
  - Frequently much better performance than *un-adapted* system (most out-of-the-box non-trainable tools)
  - Engineering requirements are relatively few
  - Manual annotation burden is relatively slight
- **Missing: statistical tools that are end user-trainable**
  - Annotation and tool configuration remain specialized, though teachable, skills

CIIS
Center for Integrated Intelligence Systems

CENTER FOR ENTERPRISE MODERNIZATION

MITRE