JRC
EUROPEAN COMMISSION

UNIVERSITY OF HELSINKI

Combining Information Retrieval and Information Extraction for Medical Intelligence

MMDSS'2007

Gazzada, Italy, 20 September 2007

Roman Yangarber, Ralf Steinberger,
Clive Best, Peter von Etter, Flavio Fuart & David Horby

http://medusa.jrc.it/
http://doremi.cs.helsinki.fi/jrc/

MedISys
Medical Information System

Outline

- Introduction: Information and Intelligence
- MedISys: Information Retrieval
- PULS: Information Extraction
- MedISys/PULS Integration
- Information Aggregation
- Performance
- Current work

Users and motivation

Use case:

- Organization = ECDC, national Health Authorities/Agencies

Goal:

- We want to deliver a simple and effective *early-warning* system

Information vs. Intelligence

- Timely – no time lost from the information's initial appearance to delivery
  - everyone is 100% intelligent in hindsight
- Complete – no information missed, from any source
  - basically: high recall
- Concise – no information overload
  - not quite about precision, but related

- Last requirement – conciseness – has two complementary aspects:
  - no redundancy, no unnecessary detail:
    - deliver only the minimum necessary information
    - If the user wishes more detail, s/he can request further elaboration
  - comprehensive background knowledge:
    - for a system to identify *genuinely new* and *critical* information, with high confidence, it must know whether this information truly new, or has already been known, at some earlier time or from some different source

Combination of Technologies

Delivering **intelligence** that is critical for the organization's functioning entails three key technologies:

- Information Retrieval
- Information Extraction
- Information Aggregation

Outline

- Introduction: Information and Intelligence
- MedISys: Information Retrieval
- PULS: Information Extraction
- MedISys/PULS Integration
- Information Aggregation
- Performance
- Current work

Medical Information System - MedISys

- Functionality of MedISys:
  - Gather documents of relevance
  - Classify
  - Aggregate information
  - Alert users

- *Europe Media Monitor* (EMM) family of applications
  - NewsBrief (http://press.jrc.it/)
  - MedISys (http://medusa.jrc.it/)
  - NewsExplorer (http://press.jrc.it/NewsExplorer/)

  → Input > 1,000 news portals world-wide

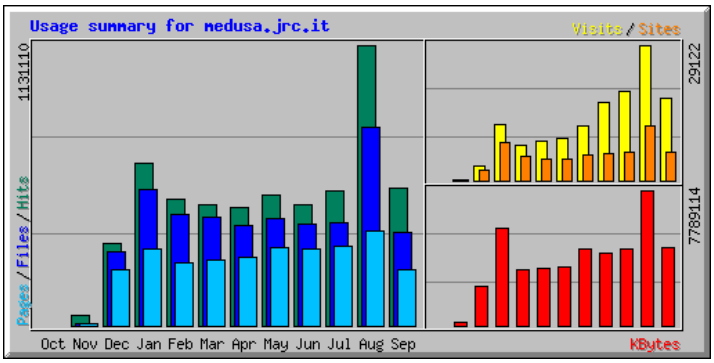
  →~ 35,000 news items per day

  →In 34 languages

  \+ 150 medical web sites + more local newspapers

Public vs. restricted MedISys

- Public MedISys site, since 2007: http://medusa.jrc.it/,
  - http://medusa.jrc.it/medisys/alertedition/diseases/all/Ebola.html
  - Currently ~ 1,000 visits / day



- Restricted MedISys site additionally
  - Receives newswires
  - Receives pay-for documents (Lexis Nexis)
  - Offers more categories
  - Offers additional functionality (daily summary reports, SMS alert, …)

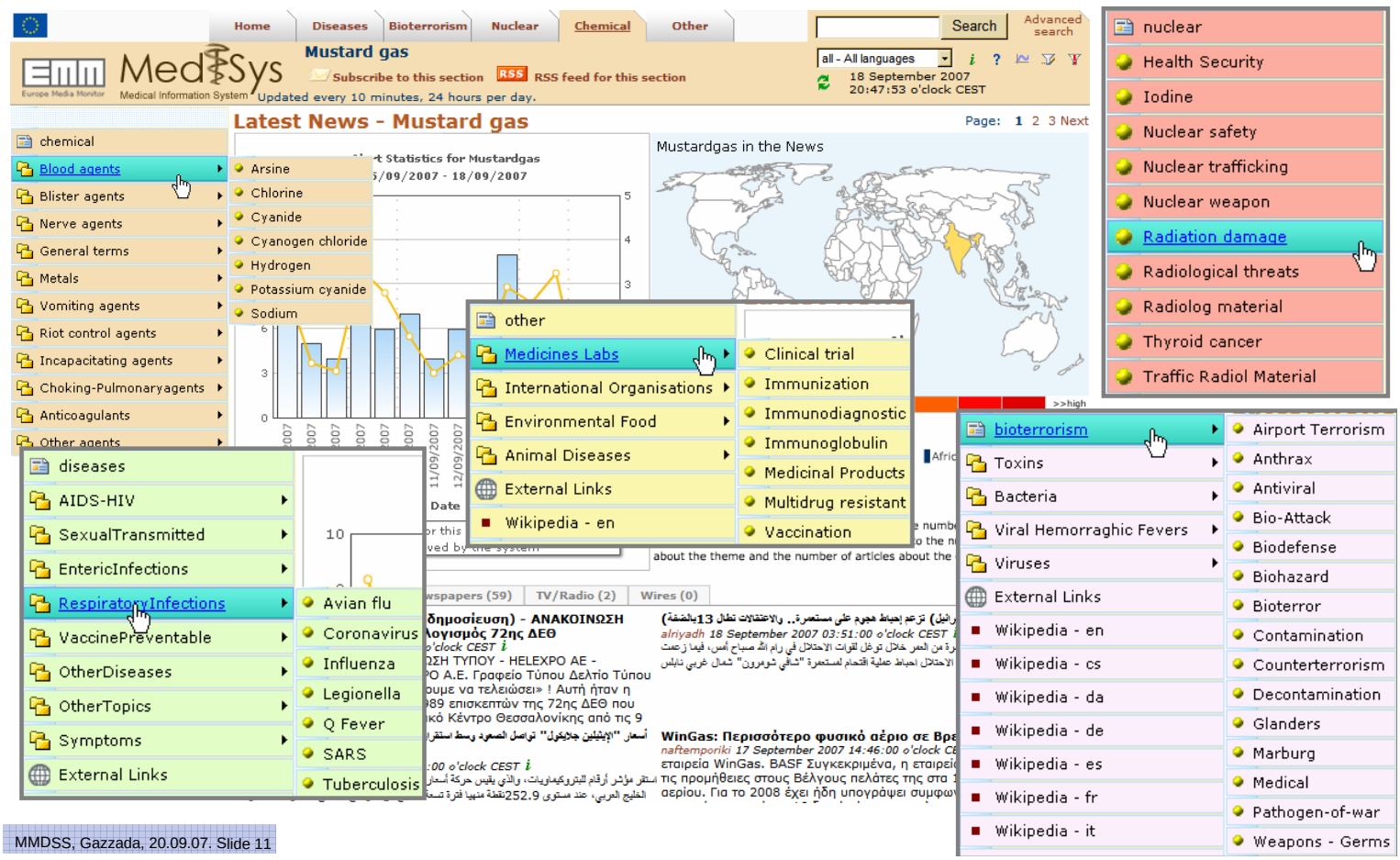
MedISys - Objective

- Initiated by European Commission Directorate General 'Health and Consumer Protection'
- Objective:
  - Support national and international Public Health (PH) organisations
    - Gather documents of relevance
    - Filter and classify
    - Aggregate information
    - Alert users
  - Replace manual scanning of multiple newspapers and web sites
  - To monitor issues of Public Health concern:
    - Outbreak of contagious diseases
    - Nuclear and chemical incidents
    - Bioterrorism, harmful substances (anthrax, crowd control agents, …)
    - Info on Public Health organisations themselves
    - …

Current Subscribers to MedISys Alerts and Reports include

- European Centre for Disease Control, Stockholm
- World Health Organisation
- Euro-Surveillance
- Swiss Federal Office of Public Health
- Icelandic Ministry of Health
- Spanish Ministry of Sanitation
- Spanish Ministry of Health and Consumer Protection
- Institut de Veille Sanitaire France
- Global Public Health Intelligence Network (Canada)
- Danish Emergency Management Agency
- Italian Ministry of Health
- Italian Ministry of Defence
- Dutch Institute of Public Health
- Dutch Food and Consumer Product Safety Authority

MedISys categories and category types



Filtering of Public Health-related news

~ 260 different categories ('alerts') + one for each country

- Boolean search word expressions
- Cumulative weights
- Vicinity
- Wild cards
- Multilingual (example: avian flu)

A combination of at least one of

grip_
grippe
gryp
influenz%
flu
gripa%
gripe
gripă
influenca
kuře
letadlo
madár
pasăre
pui
الطيري
الطائر
chrípka
chřipka
гриппа%

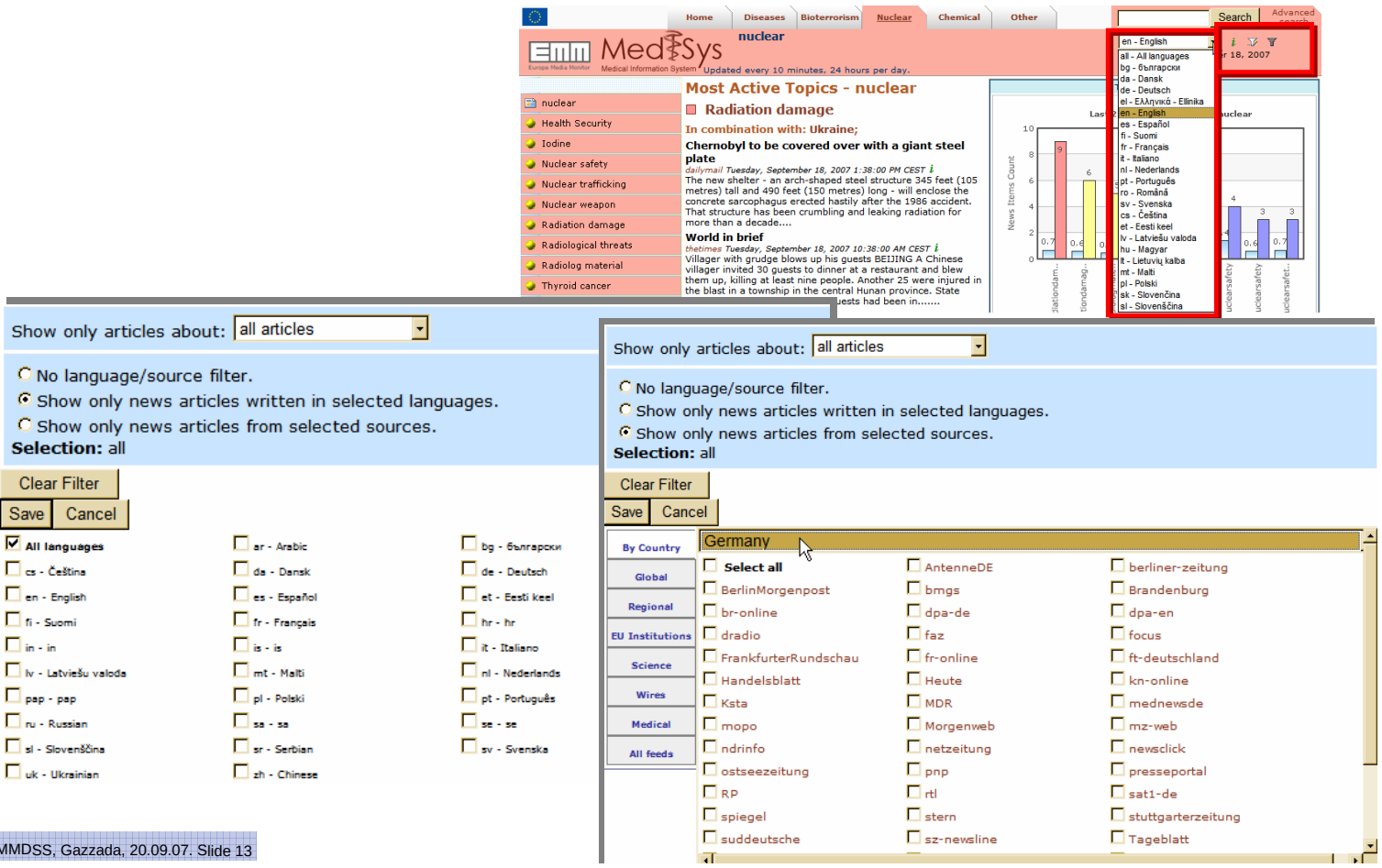
and at least one of

aviaria
aviária
ptasi%
птич%
aviair%
avian
aviar
bird
chicken
csirke
tič
vtáčí
paukščių
putka
kurča
putnu
الإنفلونزا

Pattern	Weight
الطيور+	20
fjerkræinfluenza	20
h5n1	20
h2n2	20
h3n2	20
fugleinfluenza	20
geflügelpest	20
linnugripp	20
lintuinfluenssa	20
vogelgriep	20
vogelgrippe	20
vogelpest	20
أنفلونزا	20
鸟流感	20
γρίπη+των+πτηνών	20
γρίπη+πουλερικών	20

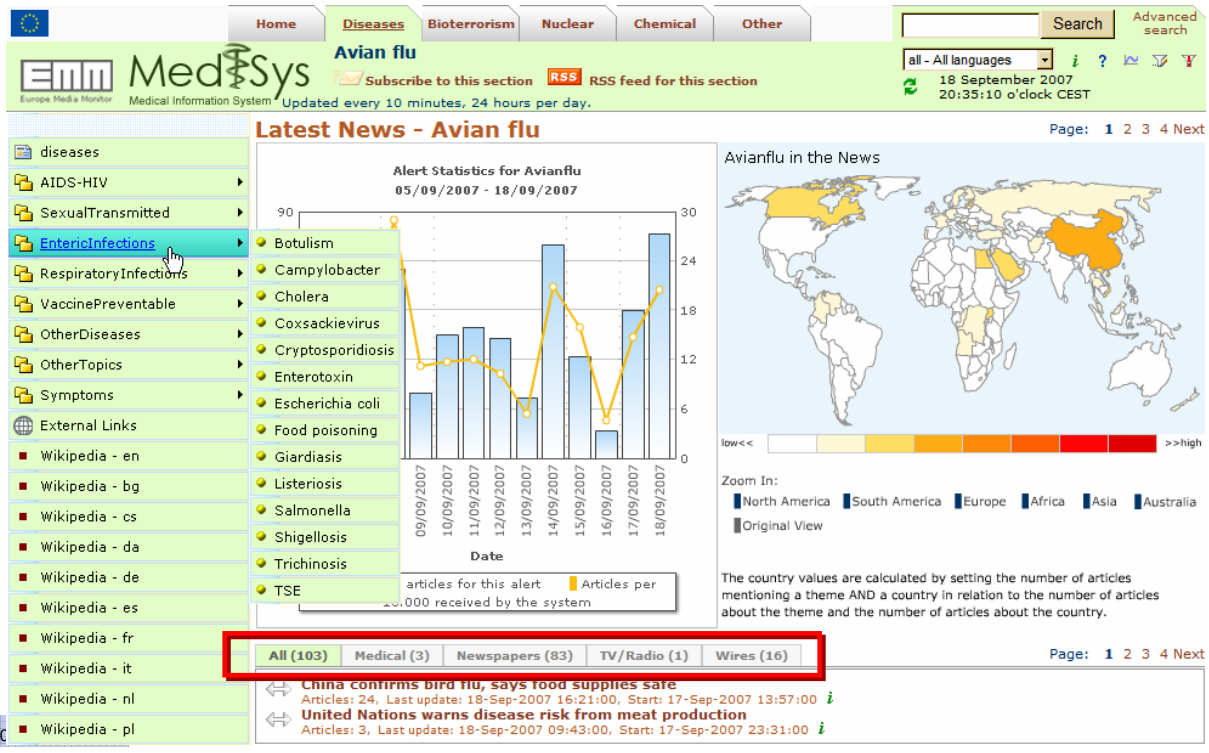
Threshold: 20

Filtering news by language and sources



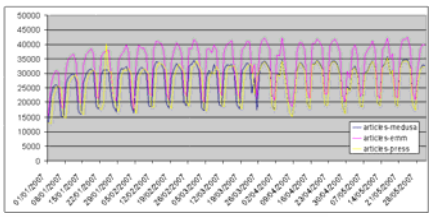
Aggregation of the multilingual 'alert' statistics (1)

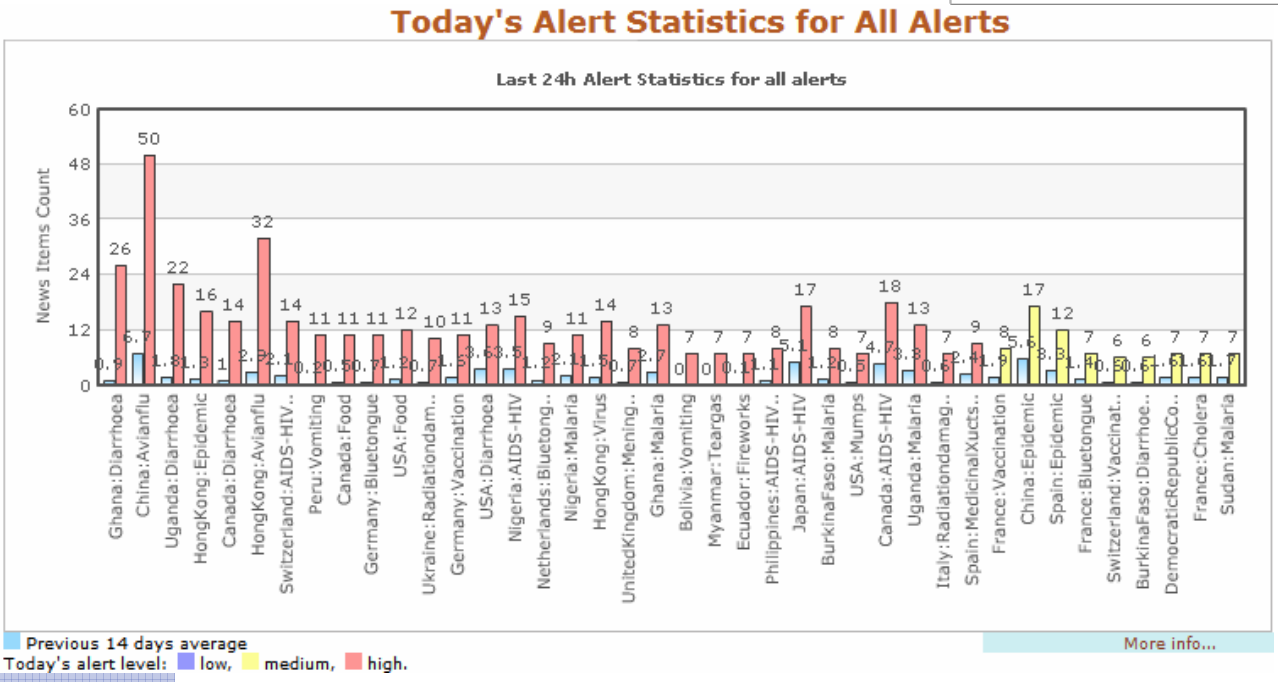
- Intersection of medical category and country
- Shown summary statistics are based on aggregated information from all 34 languages



Aggregation of the multilingual 'alert' statistics (2)

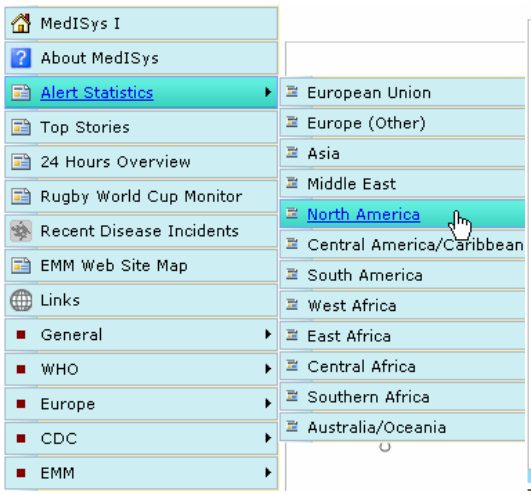
- Calculate alert level by comparing last 24 hours with 2-week average
- Normalisation of variation of number of daily articles
- Assuming normal distribution of articles per category per day, using multiples of stdev to determine alert level

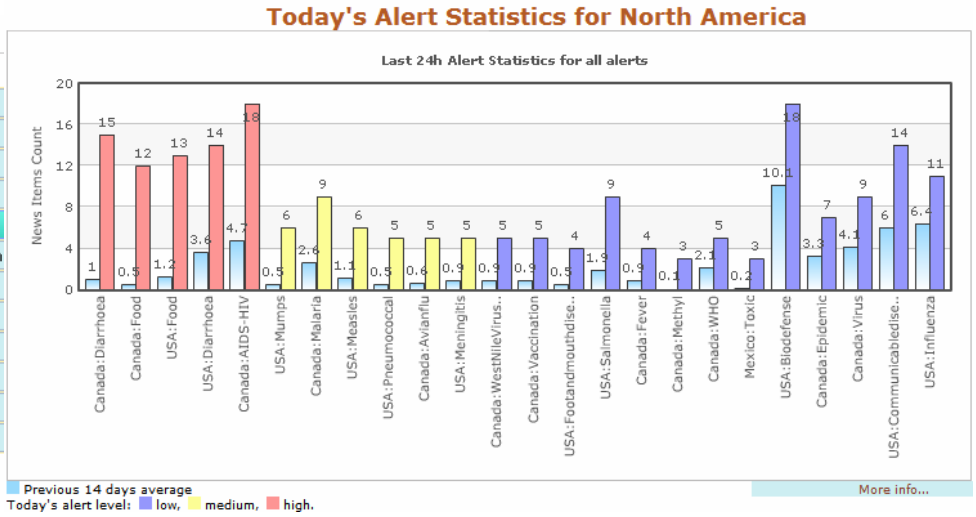




Aggregation of the multilingual 'alert' statistics (3)

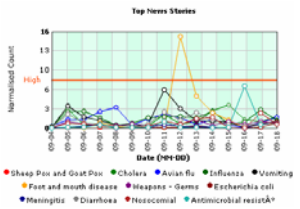
- Alert statistics by region

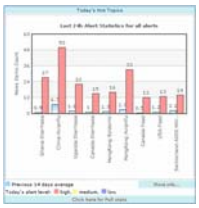


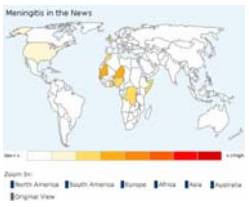


Alerting functions

- Various online statistics



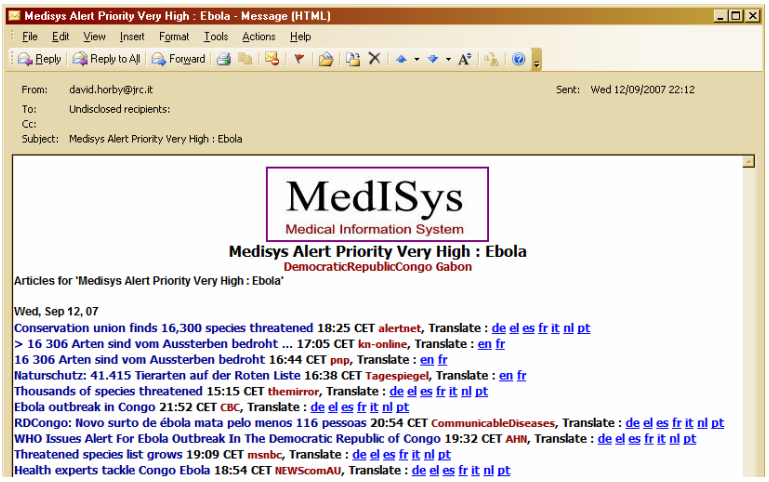




- Email subscriptions
  - Immediate
  - Daily

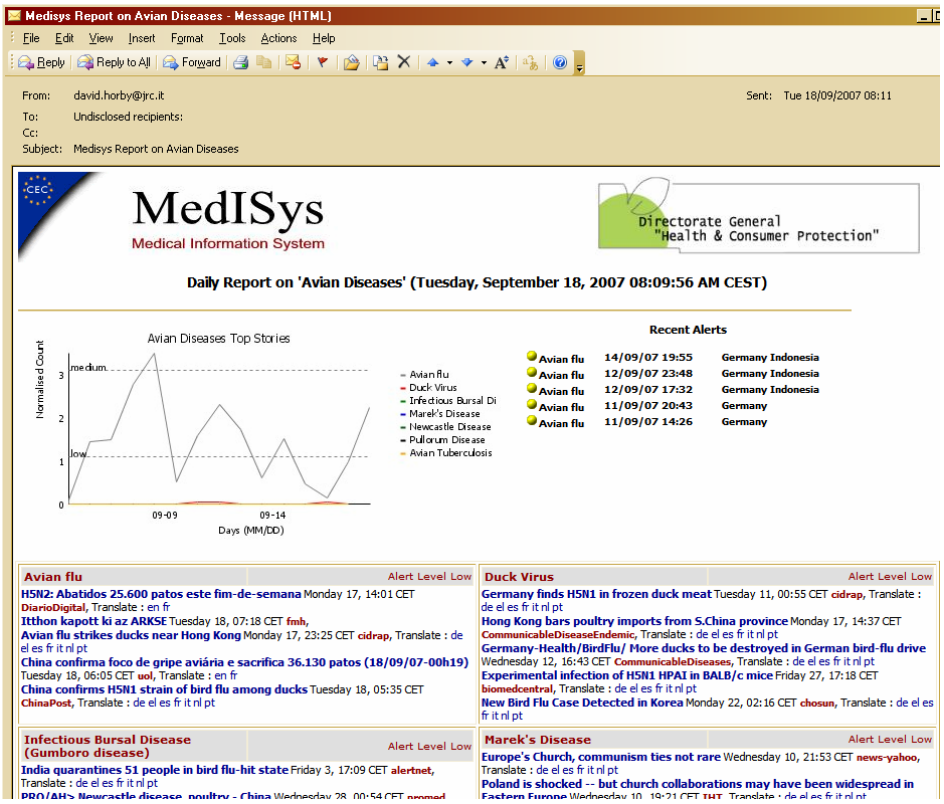
  - Alert level-dependent

Enteric Infections				
Cholera	○ Low	○ Medium	○ High	◉ None
Escherichia coli	○ Low	○ Medium	◉ High	○ None
Salmonella	○ Low	○ Medium	◉ High	○ None
Shigellosis	○ Low	○ Medium	◉ High	○ None
Listeriosis	○ Low	○ Medium	○ High	◉ None
Botulism	○ Low	○ Medium	○ High	◉ None



Alerting functions (2)

- Email subscriptions
  - Summary reports



- SMS alerts

Medical events in MedISys

- Information Retrieval Statistics are based on *mentions* of disease names
  - Triggered by news on outbreaks, discussions about vaccinations, new medicines, summary reports, …

→ Use Information Extraction to specifically identify *incidents*, or *outbreaks*

→ *English* articles found on websites are fed to medical event extraction system PULS

Outline

- Introduction: Information and Intelligence
- MedISys: Information Retrieval
- PULS: Information Extraction
- MedISys/PULS Integration
- Information Aggregation
- Performance
- Current work

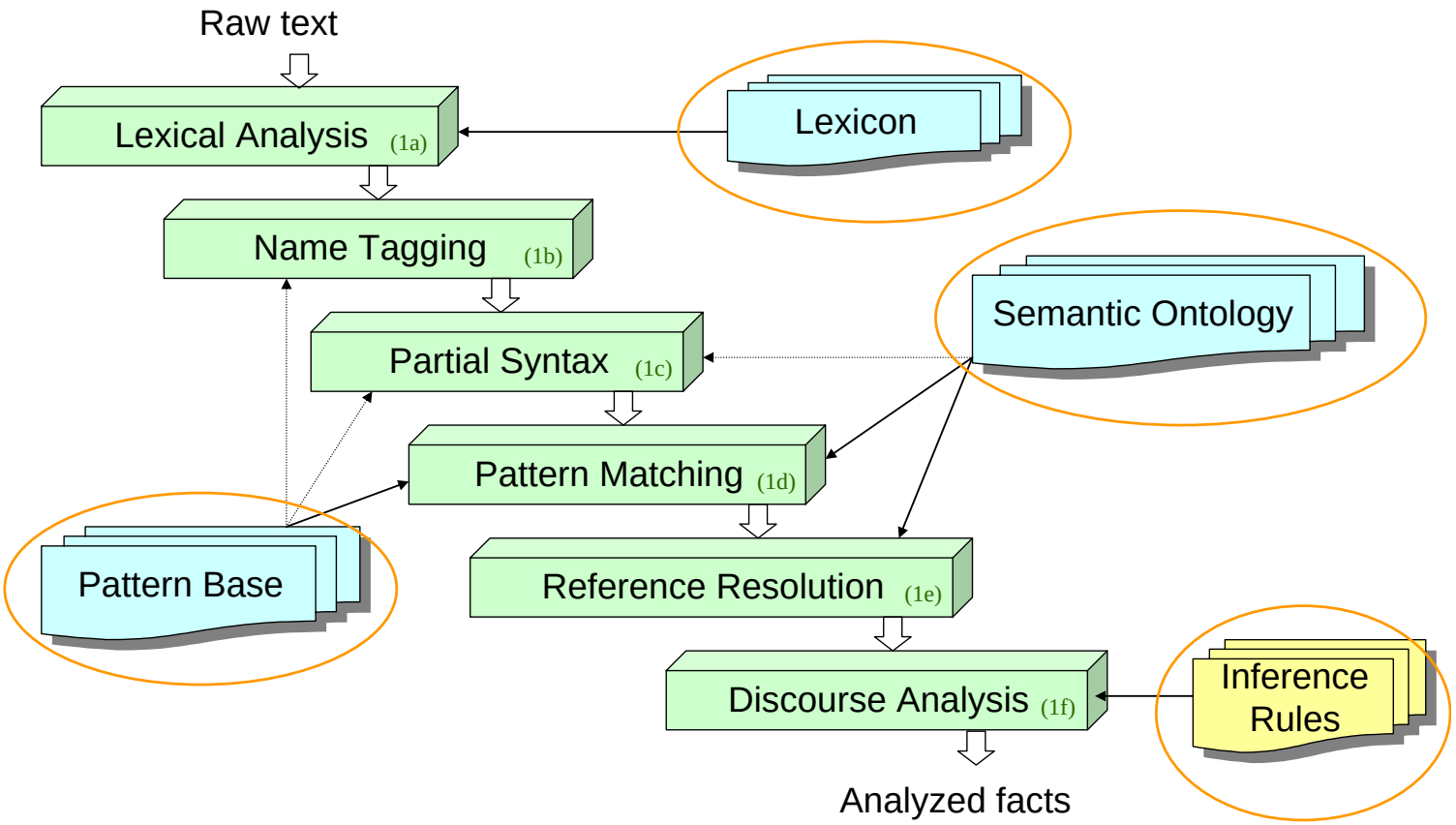
MedISys → Beyond IR

Rationale: IE can further enhance functionality of MedISys

- Provide specific facts, extracted from the documents found by MedISys
- Boost precision
  - keyword-based queries may trigger on documents which are off-topic but happen to mention the alerts in unrelated contexts
  - pattern matching in IE provides the mechanism that assures that the keywords appear in relevant contexts only

Event Extraction review

Core IE Engine and Knowledge bases (KBs)



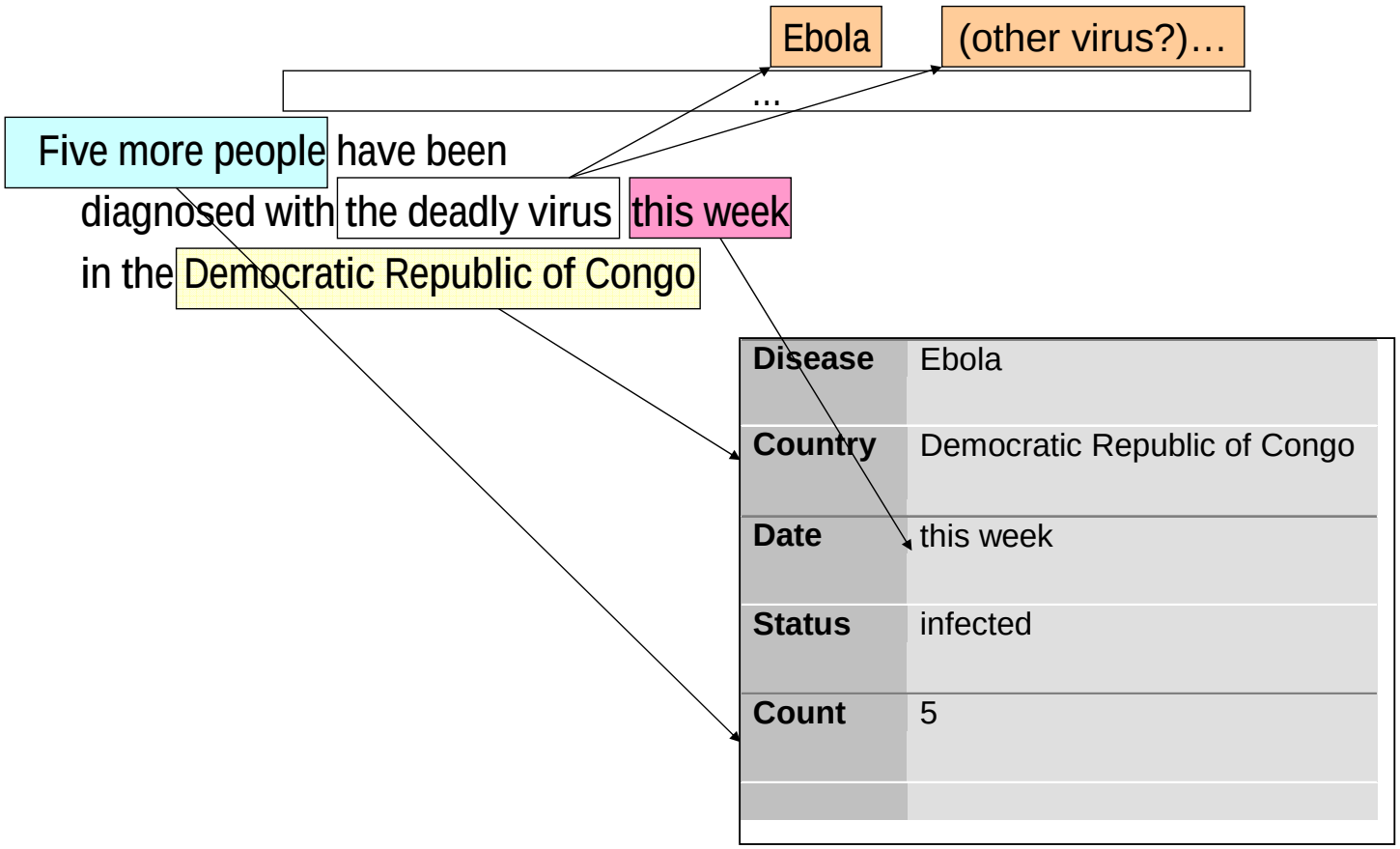
Example: Event extraction

np(Victim) *vp-pass(diagnose)* *'with'* *np(Disease)* [*np(Date)*] [*'in'* *np(Location)*]

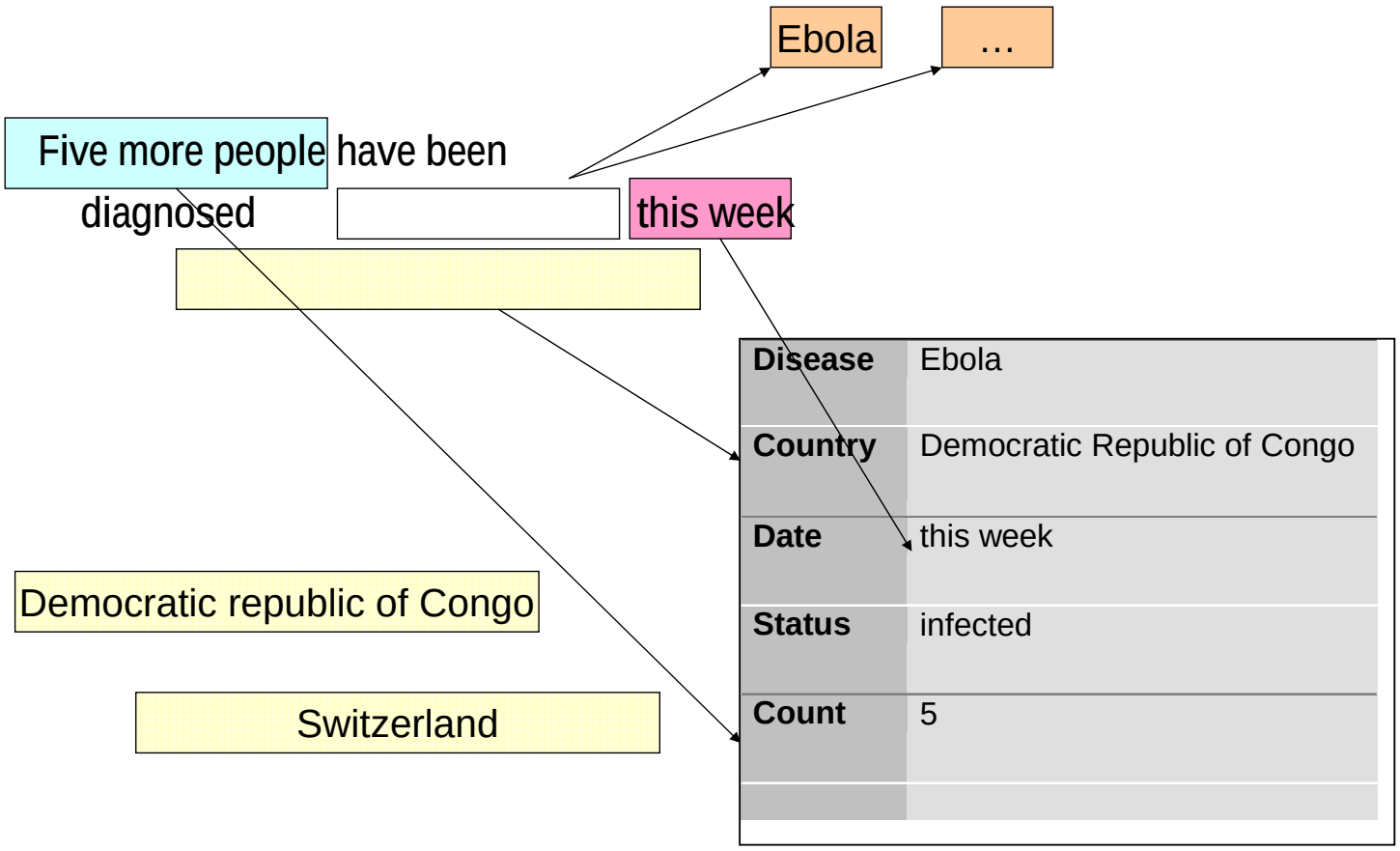
Five more people have been diagnosed with Ebola during this week in the Democratic Republic of Congo

Disease	Ebola
Country	Democratic Republic of Congo
Date	this week
Status	infected
Count	5

IE and Semantics: Reference resolution

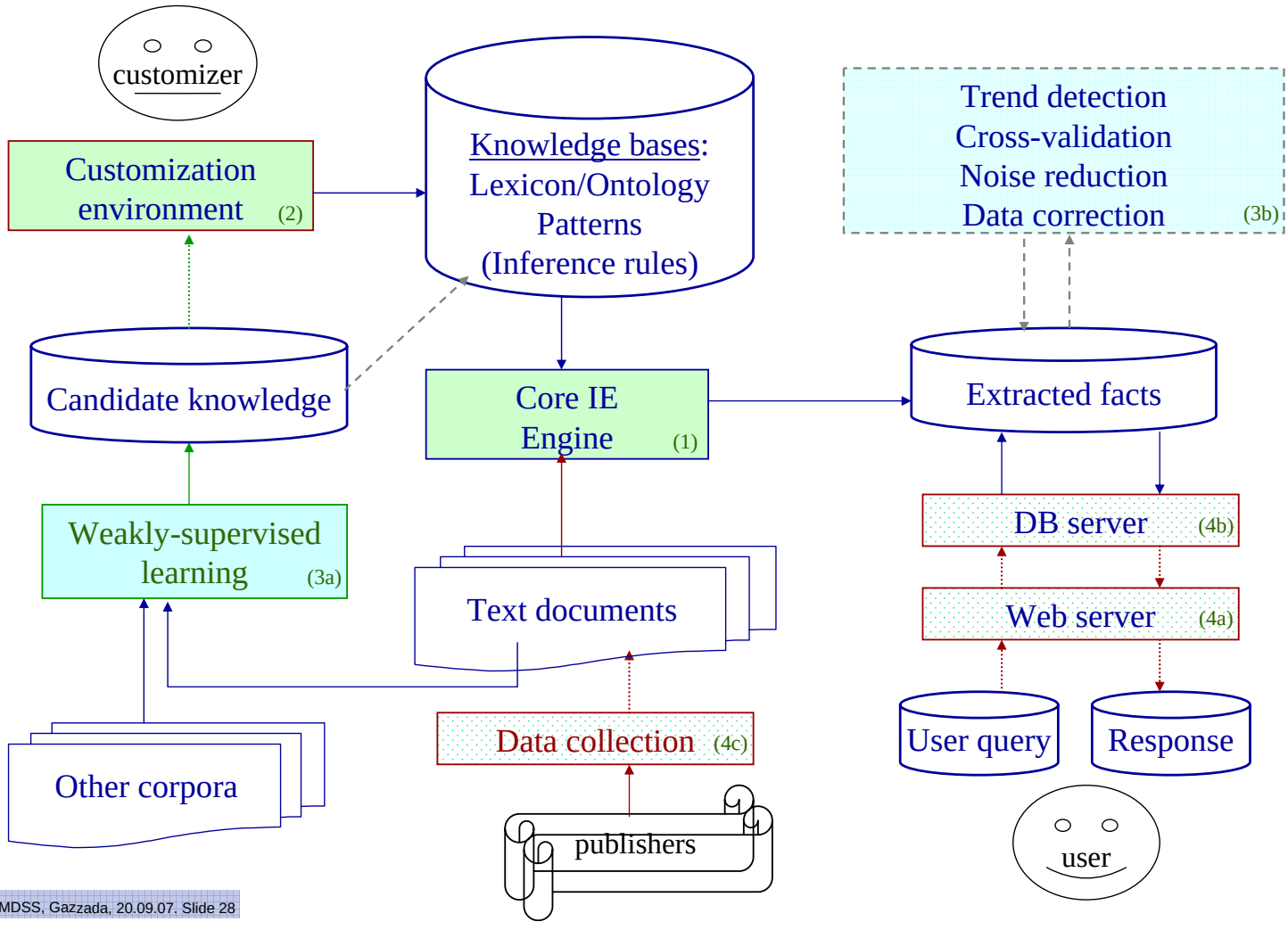


IE and Semantics: Elided attributes



PULS

- Pattern-based Understanding and Learning System
- Developed at University of Helsinki, Finland
- Customized for several application domains: *epidemic surveillance*

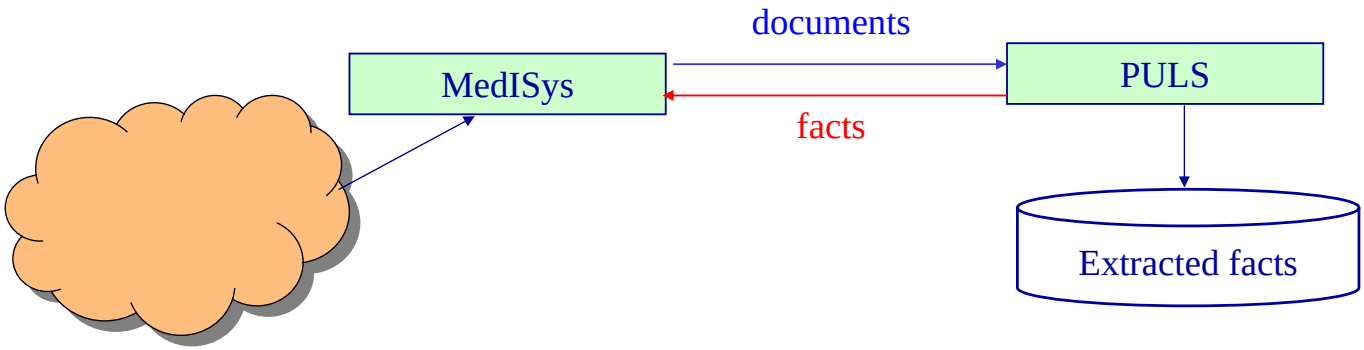


Outline

- Introduction: Information and Intelligence
- MedISys: Information Retrieval
- PULS: Information Extraction
- MedISys/PULS Integration
- Information Aggregation
- Performance
- Current work

MedISys/PULS Integration

- RSS channel
- MedISys sends new documents
- PULS returns new events
- Live exchange of data every 10 minutes



- Demo: records that PULS returns to MedISys
  - top five: latest, most urgent
  - top fifty

MedISys + PULS

The Whole vs. the sum of its parts?

- Large volume of information
  - and growing: began collecting data this year
- Continuous stream of documents and events

- Organizing this information?
  - Need
  - Opportunity

Outline

- Introduction: Information and Intelligence
- MedISys: Information Retrieval
- PULS: Information Extraction
- MedISys/PULS Integration
- Information Aggregation
  - Local confidence
  - Global grouping
- Performance
- Current work

Toward Cross-Document Aggregation

Traditional IE paradigm:

1. Documents are processed *independently*:
   - facts found in one document do not interact with information found in other documents
   - rationale: inexact!
2. For each attribute, the system stores only a *single, "best"* value:
   the best guess – based on local features

Two steps beyond the traditional paradigm:

- after extracting information locally, attempt to unify it *globally*:
  - in this domain: generate "outbreaks"
- when we store a record in a database, do not store a single value,
  - store other, less probable values as well

Distribution of attribute values

In general, each attribute will have *several* candidate values in the document

- attach a score to each candidate, based on local, heuristic features
  - how well does this candidate fit the event?
  - the value is mentioned inside the *trigger*
  - appears in the same sentence as the trigger
  - whether it appears before or after the sentence containing the trigger
  - the candidate appears in the document's headline
  - the number of times the candidate appears in the text
  - this value is the unique value of its type in the sentence
  - ...
- The scores give the distribution
- Use set of candidate values rather than a single candidate
  - it allows us to compute the **confidence** of an incident, which is used in cross-document aggregation
  - methods for recovery from locally-best but incorrect guesses by using global information.

Confidence

Compute confidence for each incident

- Local heuristics
- Based on confidences of the attributes
  - A "core" set of principal attributes:
  - Disease name
  - Location (State)
  - Date

- If score of attribute > θ → confident attribute
- If all principal attributes confident → confident tuple

Utilizing Confidence

- Potentially useful in itself:
  - user can restrict search by confidence (we will show how effectivethis is)
- "Lend" confidence to other records

Aggregation into Outbreaks

Simple idea:

- Outbreak = chain of incidents/records
- Connection criteria:
  - "related" disease
  - "nearby" in location
  - "nearby" in time

- Somewhat simplistic: in reality outbreaks are graphs of incidents
  - 10 people contracted the virus in Uganda since the start of the year
  - 5 health workers died in Gulu province last month

NB: Aggregation takes place across incidents:
- documents
- sources

Outline

- Introduction: Information and Intelligence
- MedISys: Information Retrieval
- PULS: Information Extraction
- MedISys/PULS Integration
- Information Aggregation
- Performance
- Current work

Performance: some preliminary numbers

Averages since April 2007:

	MedISys documents	PULS incidents	
produce event	2700	6000	
produce no event	7300	—	**false negative 15%** (of 100)
Total	10000		

From MedISys perspective:

- ~62% documents contained no incident (or more)
- Good news: PULS helps distinguish epidemics from other discussions

Evaluation of Confidence

- System rates 20% of all incidents as confident
- Manually checked 100 random incidents for accuracy of confidence heuristics:
- Very coarse (pessimistic) judgements: no partial credit

correct	**72%**
incorrect (one principal attribute wrong)	**14%**
erroneous (non-event)	**14%**

Optimistic judgement:

- Incorrect: not much hope: inherent complexity in the text
- Erroneous: could fix with some tuning labor: refine patterns
- Max. 83.7% correct

Evaluation of Outbreak Aggregation

Important to evaluate

- Primary means of reducing redundancy in flow of news
- Manually analyzed 20 outbreaks of about 10 incidents each

Measures:

- 68% of the incidents were correctly identified with their outbreaks
  - including non-confident incidents
  - three of the outbreaks (15%) were erroneous
    - built upon incorrect confident incidents

- Only 22.5% of the examined incidents were confident
  - on average, outbreaks contained 2-3 confident incidents
- Aggregation is often useful even when the outbreak consists entirely of incorrectly analyzed incidents:
  - Lump together similar reports

Outline

- Introduction: Information and Intelligence
- MedISys: Information Retrieval
- PULS: Information Extraction
- MedISys/PULS Integration
- Information Aggregation
- Performance
- Current work

Improvements

- Event detection in multiple languages

- Global confidence, aggregated across documents/sources
- Combining sources of evidence
- Agreement between MedISys "outbreaks" and PULS "outbreaks“
  - Urgency:
  - MedISys urgency criterion: something is newly prominent across news sources
    - by the time it's prominent, potentially already dated
  - Incorporating PULS confidence may help promote events to prominence sooner

  → early-warning system