

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/>



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

- 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

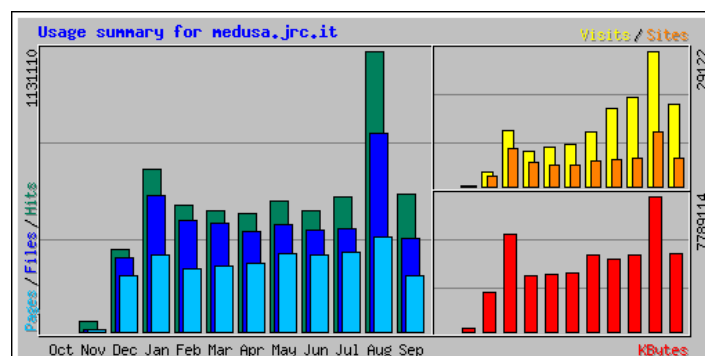
→ In 34 languages

+ 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, ...)

- 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



MedSys categories and category types

The screenshot displays the MedSys website interface. At the top, there are navigation tabs for Home, Diseases, Bioterrorism, Nuclear, Chemical, and Other. A search bar and advanced search options are also present. The main content area features a 'Latest News - Mustard gas' section with a bar chart and a world map. On the left, there are several vertical navigation menus: 'chemical' (with sub-items like Blood agents, Blister agents, etc.), 'diseases' (with sub-items like AIDS-HIV, Respiratory Infections, etc.), and 'bioterrorism' (with sub-items like Toxins, Bacteria, etc.). A central pop-up menu is open, showing 'other' categories like Medicines Labs, International Organisations, etc. On the right, a sidebar menu lists various categories such as nuclear, Health Security, Iodine, etc.

MMDSS, Gazzada, 20.09.07. Slide 11

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 and at least one of

grip_
grippe
gryp
influenz%
flu
gripa%
gripe
gripã
influenca
kuře
letadlo
madár
pasãre
pui
الطيري
العاثر
chřipka
chřipka
грипп%

aviaria
aviãria
ptasi%
птич%
aviair%
avian
aviar
bird
chicken
csirke
tič
vtãči
paukščių
putka
kurča
putnu
الإنطونزا

Pattern	Weight
الطير+ +الطير	20
fjerkraeinfluenza	20
h5n1	20
h2n2	20
h3n2	20
fugleinfluenza	20
gefługapest	20
linnugripp	20
lintuinfluenssa	20
vogelgriep	20
vogelgrippe	20
vogelpest	20
أنطونزا	20
鳥流感	20
γριππη+των+πτηνων	20
γριππη+πουλερικων	20

Threshold: 20

MMDSS, Gazzada, 20.09.07. Slide 12

Filtering news by language and sources

Show only articles about:

No language/source filter.
 Show only news articles written in selected languages.
 Show only news articles from selected sources.
Selection: all

Clear Filter

Save Cancel

<input checked="" type="checkbox"/> All languages	<input type="checkbox"/> ar - Arabic	<input type="checkbox"/> bg - Български
<input type="checkbox"/> cs - Čeština	<input type="checkbox"/> da - Dansk	<input type="checkbox"/> de - Deutsch
<input type="checkbox"/> en - English	<input type="checkbox"/> es - Español	<input type="checkbox"/> et - Eesti keel
<input type="checkbox"/> fi - Suomi	<input type="checkbox"/> fr - Français	<input type="checkbox"/> hr - Hrvatski
<input type="checkbox"/> in - In	<input type="checkbox"/> is - Íslenska	<input type="checkbox"/> it - Italiano
<input type="checkbox"/> lv - Latviešu valoda	<input type="checkbox"/> mt - Malti	<input type="checkbox"/> nl - Nederlands
<input type="checkbox"/> pap - Papiamentu	<input type="checkbox"/> pl - Polski	<input type="checkbox"/> pt - Português
<input type="checkbox"/> ru - Russian	<input type="checkbox"/> sa - Sanskrit	<input type="checkbox"/> se - Svenska
<input type="checkbox"/> sl - Slovenščina	<input type="checkbox"/> sr - Srpski	
<input type="checkbox"/> uk - Ukrainian	<input type="checkbox"/> zh - Chinese	

Show only articles about:

No language/source filter.
 Show only news articles written in selected languages.
 Show only news articles from selected sources.
Selection: all

Clear Filter

Save Cancel

Germany

By Country: Select all

<input type="checkbox"/> Global	<input type="checkbox"/> AntenneDE	<input type="checkbox"/> berliner-zeitung
<input type="checkbox"/> Regional	<input type="checkbox"/> bmgs	<input type="checkbox"/> Brandenburg
<input type="checkbox"/> EU Institutions	<input type="checkbox"/> dpa-de	<input type="checkbox"/> dpa-en
<input type="checkbox"/> Science	<input type="checkbox"/> faz	<input type="checkbox"/> focus
<input type="checkbox"/> Wires	<input type="checkbox"/> fr-online	<input type="checkbox"/> ft-deutschland
<input type="checkbox"/> Medical	<input type="checkbox"/> Heute	<input type="checkbox"/> kn-online
<input type="checkbox"/> All feeds	<input type="checkbox"/> MDR	<input type="checkbox"/> mednewsde
<input type="checkbox"/> RP	<input type="checkbox"/> Morgenweb	<input type="checkbox"/> mz-web
<input type="checkbox"/> spiegel	<input type="checkbox"/> netzeitung	<input type="checkbox"/> newsclick
<input type="checkbox"/> suddeutsche	<input type="checkbox"/> ostseezeitung	<input type="checkbox"/> presseportal
	<input type="checkbox"/> pnp	<input type="checkbox"/> sat1-de
	<input type="checkbox"/> rti	<input type="checkbox"/> stuttgarterzeitung
	<input type="checkbox"/> stern	<input type="checkbox"/> Tageblatt
	<input type="checkbox"/> sz-newsline	

MMDSS, Gazzada, 20.09.07. Slide 13

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

Alert Statistics for Avianflu
05/09/2007 - 18/09/2007

Date	Articles per day
09/09/2007	10
10/09/2007	12
11/09/2007	15
12/09/2007	18
13/09/2007	10
14/09/2007	25
15/09/2007	15
16/09/2007	10
17/09/2007	20
18/09/2007	28

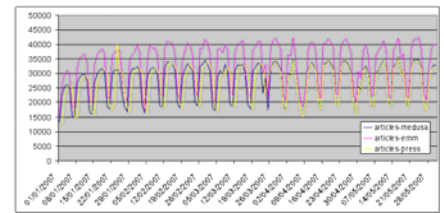
Articles for this alert: 103
Articles received by the system: 103

Summary: **All (103)** | Medical (3) | Newspapers (83) | TV/Radio (1) | Wires (16)

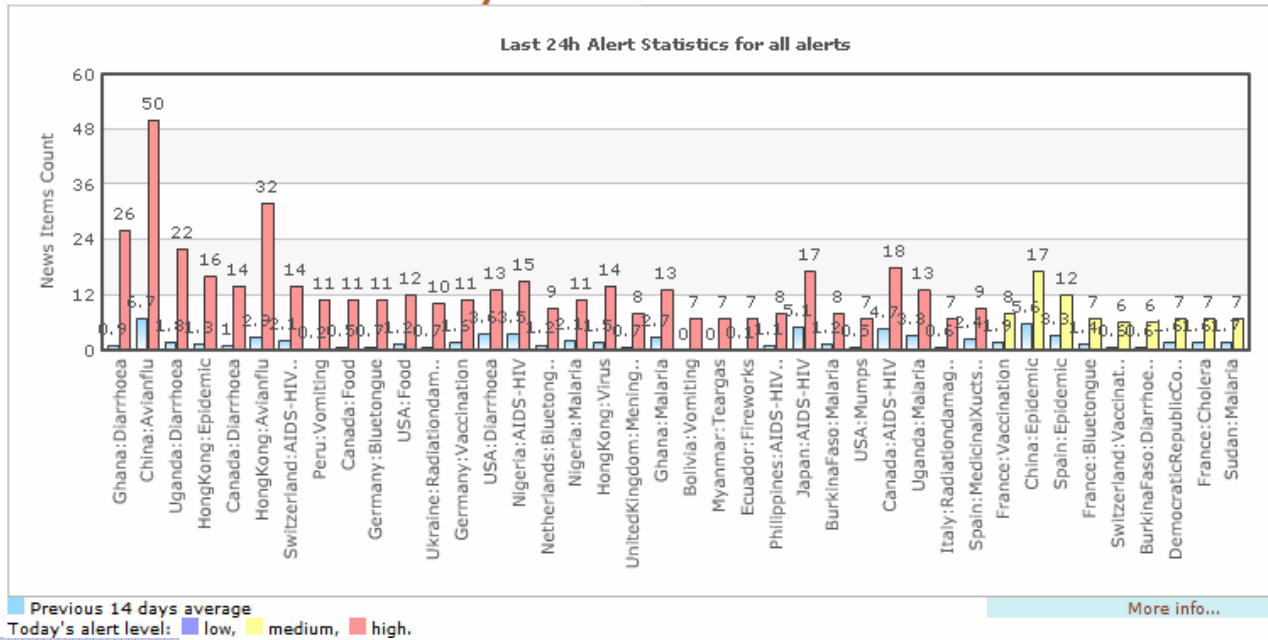
MMDSS, Gazzada, 20.09.07. Slide 14

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



Today's Alert Statistics for All Alerts

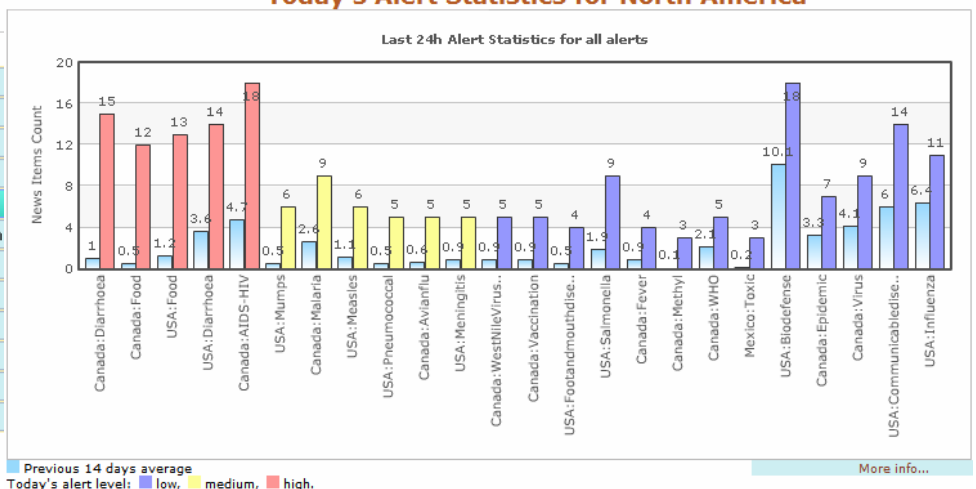


Aggregation of the multilingual 'alert' statistics (3)

- Alert statistics by region

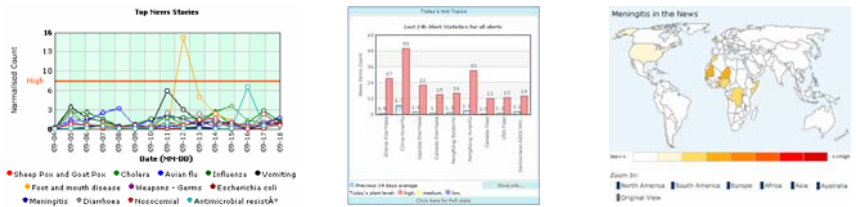
- MedISys I
- About MedISys
- Alert Statistics**
 - European Union
 - Europe (Other)
 - Asia
 - Middle East
 - North America**
 - Central America/Caribbean
 - South America
 - West Africa
 - East Africa
 - Central Africa
 - Southern Africa
 - Australia/Oceania
- Top Stories
- 24 Hours Overview
- Rugby World Cup Monitor
- Recent Disease Incidents
- EMM Web Site Map
- Links
 - General
 - WHO
 - Europe
 - CDC
 - EMM

Today's Alert Statistics for North America



Alerting functions

- Various online statistics
- Email subscriptions
 - Immediate
 - Daily
- Alert level-dependent



Enteric Infections	
Cholera	<input type="radio"/> Low <input type="radio"/> Medium <input type="radio"/> High <input checked="" type="radio"/> None
Escherichia coli	<input type="radio"/> Low <input type="radio"/> Medium <input checked="" type="radio"/> High <input type="radio"/> None
Salmonella	<input type="radio"/> Low <input type="radio"/> Medium <input checked="" type="radio"/> High <input type="radio"/> None
Shigellosis	<input type="radio"/> Low <input type="radio"/> Medium <input checked="" type="radio"/> High <input type="radio"/> None
Listeriosis	<input type="radio"/> Low <input type="radio"/> Medium <input type="radio"/> High <input checked="" type="radio"/> None
Botulism	<input type="radio"/> Low <input type="radio"/> Medium <input type="radio"/> High <input checked="" type="radio"/> None

Alerting functions (2)

- Email subscriptions
 - Summary reports
- SMS alerts



- 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



The screenshot shows the MedISys website interface. The top navigation bar includes 'Home', 'Diseases', 'Bioterrorism', and 'Other'. A search bar is located on the right. The main content area is titled 'Recent Disease Incidents' and contains a table with the following data:

Disease	Start time	End time	Location	Status	Cases	Description
Ebola Hemorrhagic Fever	2007.09.18	2007.09.18	Democratic Republic of Congo	†	166 people	DR Congo: Canada deploys mobile laboratory to assist in Ebola outbreak
Foot-and-mouth Disease	2007.09.18	2007.09.18	UK		sheep	Scientists to investigate whether sheep have role in spreading FMD
Avian Influenza	2007.09.18	2007.09.18	China		ducks	China confirms bird flu outbreak
Hepatitis	2007.09.17	2007.09.17	USA/Hawaii		inmates	Hawaii hepatitis project to aid newly released inmates
Cholera	2007.09.16	2007.09.16	Iraq		1,150 laboratory confirmed cases	Situation report on cholera outbreak in northern Iraq, 16 Sep 2007
Cholera	2007.09.16	2007.09.16	Africa		more than one million people	Fresh Rains threaten flood-hit Africa
Influenza	2007.09.16	2007.09.16	Australia			Flu outbreak hits star Sydney horses
Leishmaniasis	2007.09.15	2007.09.15	USA/Texas		nine cases	Infectious Skin Disease Found in Texas
HIV	2007.09.09	2007.09.15	Peru		three other patients	Peru blood banks face HIV crisis

Outline

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

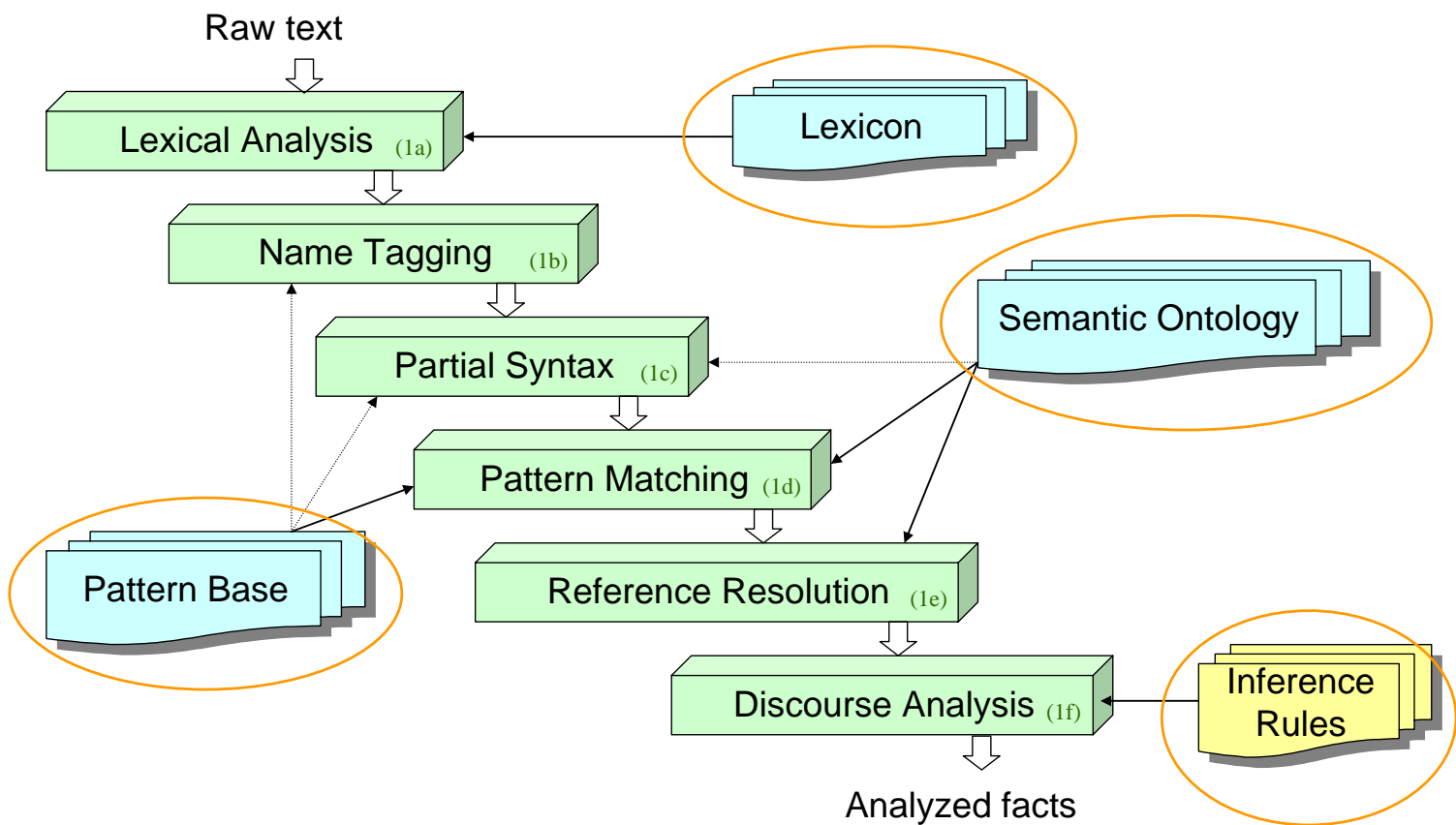
MedISys → Beyond IR

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

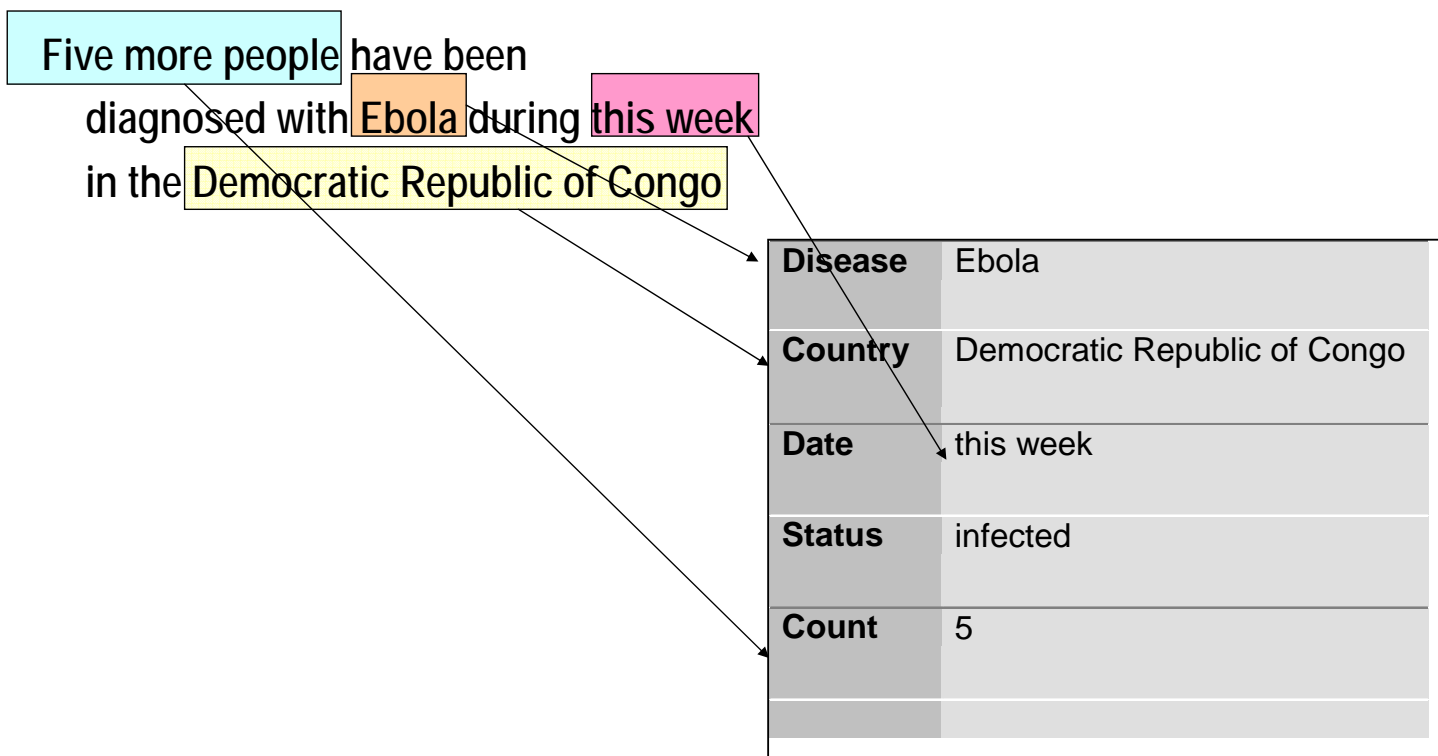
Event Extraction review

Core IE Engine and Knowledge bases (KBs)

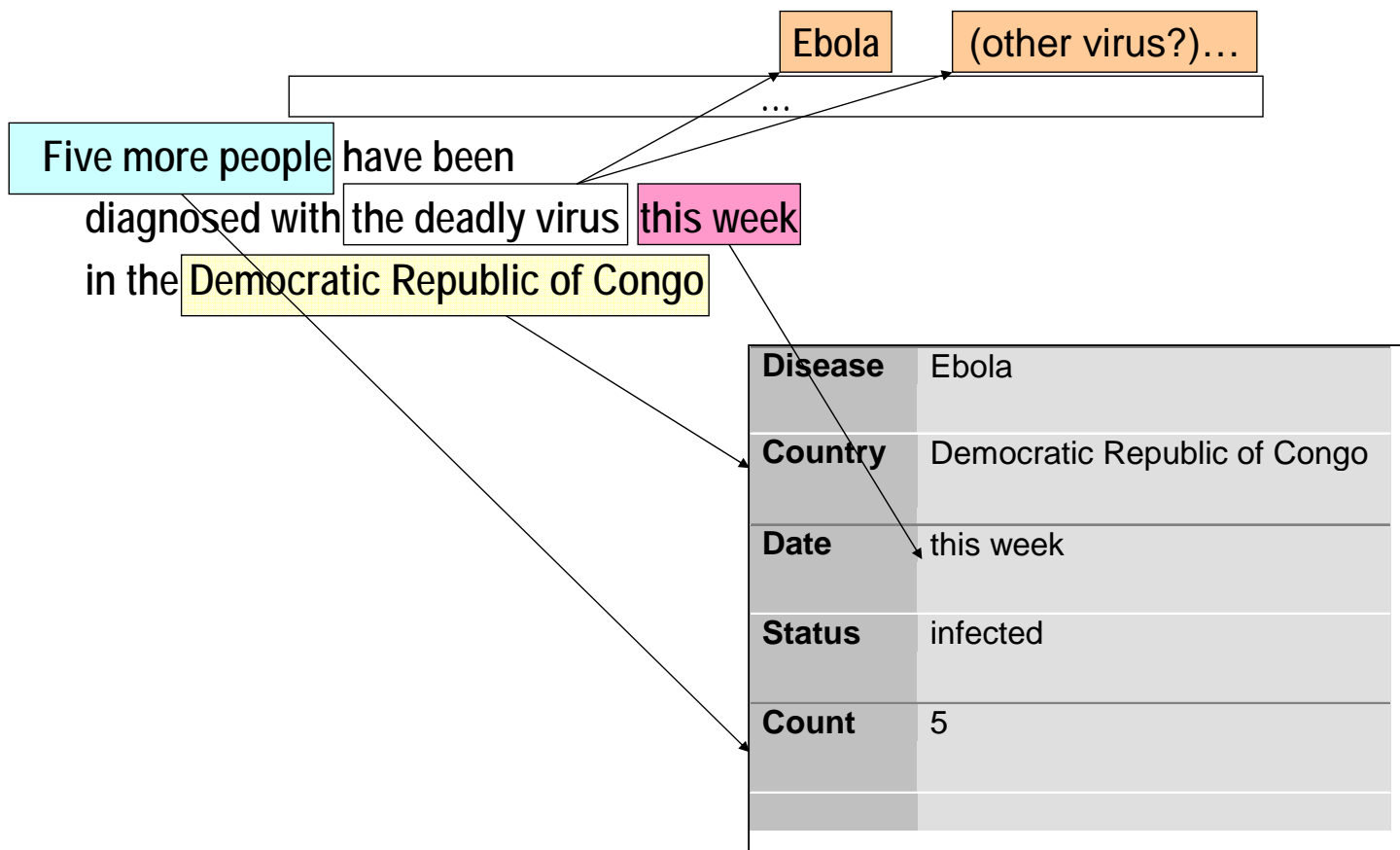


Example: Event extraction

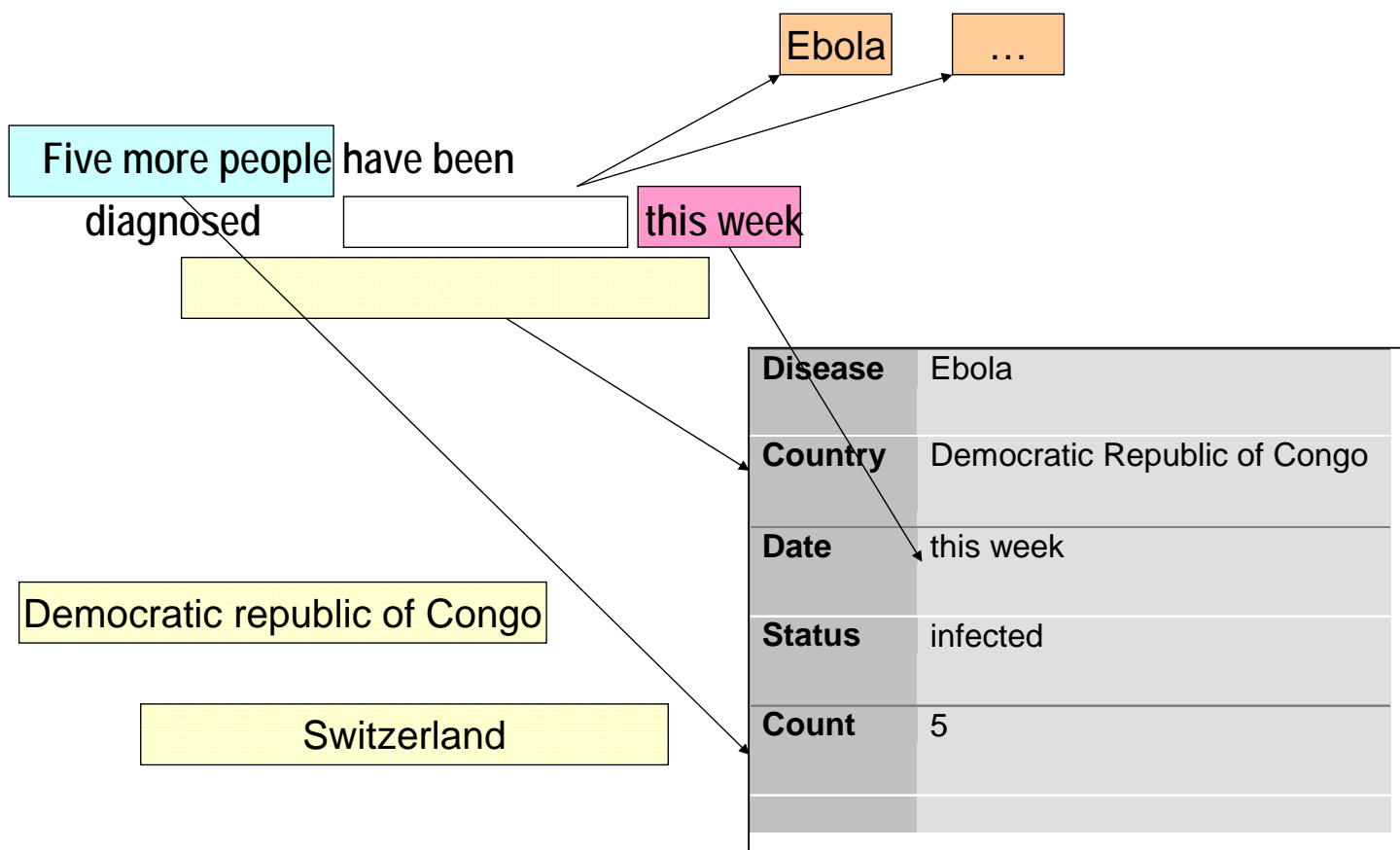
np(Victim) vp-pass(diagnose) 'with' np(Disease) [np(Date)] ['in' np(Location)]



IE and Semantics: Reference resolution



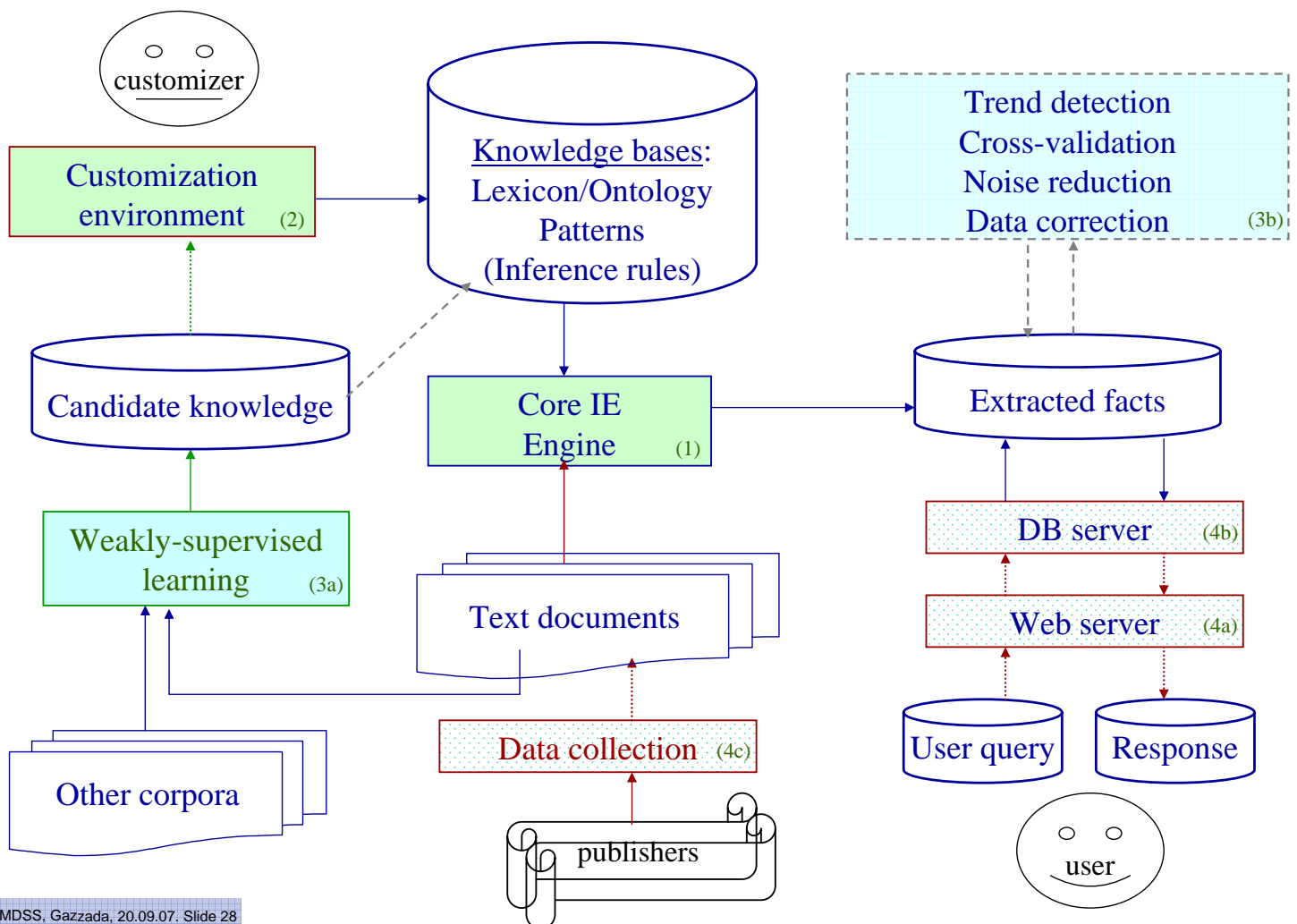
IE and Semantics: Elided attributes



PULS

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

MMDSS, Gazzada, 20.09.07. Slide 27



MMDSS, Gazzada, 20.09.07. Slide 28

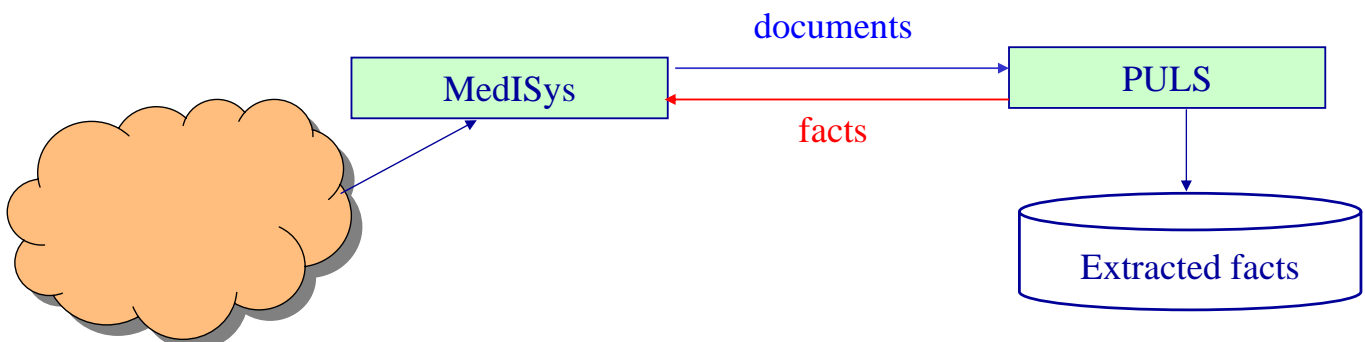
Outline

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

MMDSS, Gazzada, 20.09.07. Slide 29

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

MMDSS, Gazzada, 20.09.07. Slide 30

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 $> \theta$ → 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 effective this 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