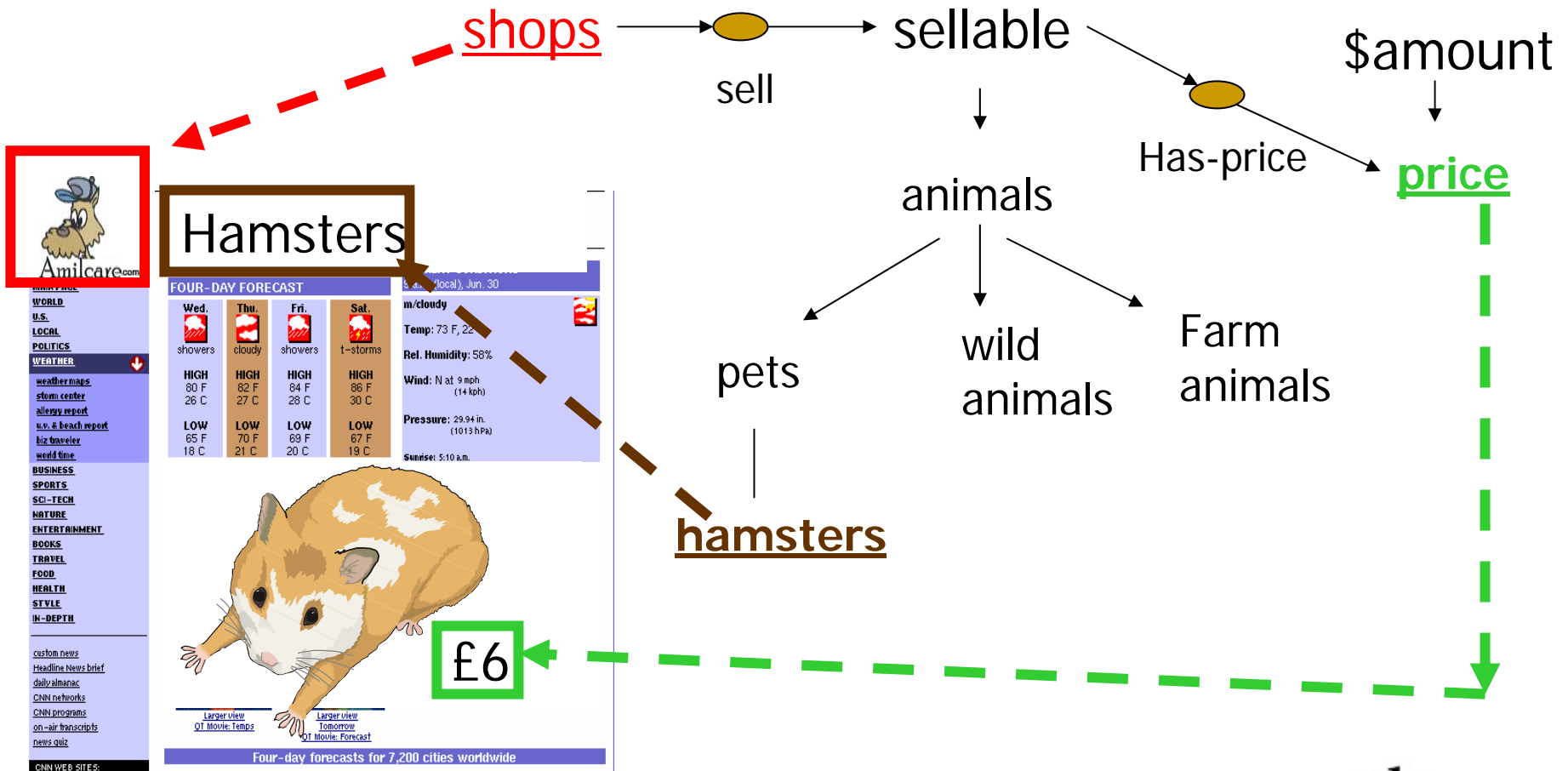

Automating Document Annotation using HLT and ML

Professor Fabio Ciravegna
Web Intelligence Technologies
Natural Language Processing Group
Department of Computer Science,
University of Sheffield

Types of Document Annotation (1)

- Marking up contained information
 - Portions of documents associated to objects in ontology
 - Enables:
 - Ontology-driven processing
 - Services based on ontology will be able to use information
 - Ontomat (Staab *et al* 2001)
 - SemTag and Seeker (Dill *et al.* 2003)
 - Armadillo (Ciravegna *et al.* 2004)
 - Melita (Ciravegna *et al.* 2002)

Ontology-based Annotation



Types of Document Annotation (2)

- Adding free text annotation (braindump)
 - The final document is just the final solution
 - Many lessons are learnt during the process and are not in the final document
 - Example: the project for a new Jet Engine
 - During the discussion the working group will consider many alternative solutions
 - Those not selected are not in the final project
 - When next jet engine is designed, the group needs to know
 - What solutions were tried (use of titanium)
 - Why they were not adopted (e.g. too high a cost)
 - If the analysis is still true (titanium cost has decreased)
 - Annotea (Barstow *et al* 2001)
 - Semantik (Gilardoni *et al* 2004)
 - AktiveDoc (Lanfranchi *et al.* 2005)

Braindump in a Legal Scenario

RICH ANDREWS

OBJECTIVE
[Click here and type objective]

EXPERIENCE

1990-1994 Arbor Shoe Southridge, SC
National Sales Manager
 == Increased sales from \$50 million to \$100 million.
 == Doubled sales per representative from \$5 million to \$10 million.
 == [redacted] products that increased earnings by 23%.

1985-1990 Ferguson and Bardell Southridge, SC
District Sales Manager
 == Increased region [redacted] million to \$350 million.
 == Managed 250 sales representatives in 10 Western states.
 == Implemented training course for new recruits — speeding profitability.

1980-1984 Duffy Vineyards Southridge, SC
Sales Representative
 == [redacted] revenues for each sales associate.
 == Expanded sales to include mass [redacted].
 == Expanded sales team from 50 to 100 representatives.

EDUCATION

1971-1975 Southridge State University Southridge, SC
 == B.A., Business Administration and Computer Science.
 == Graduated Summa Cum Laude.

INTERESTS
Southridge Board of Directors, running, gardening, carpentry, computers.

TIPS
Select text you would like to replace, and type your information.

PAX (123) 958-7454 • E-MAIL MR@MYCOMPANY.COM
12345 MAIN STREET • ANY CITY, STATE OR PROVINCE 12345-6789 • PHONE (123) 456-7890

Why we used these references

Objective [Click here and type objective]

Experience 1990-1994 Arbor Shoe Southridge, SC
National Sales Manager

- Increased [redacted] to \$100 million.
- Doubled sales per representative from \$5 million to \$10 million.
- Suggested new [redacted] products that increased earnings by 23%.

Objective [Click here and type objective]

Experience 1990-1994 [redacted] Southridge, SC
National Sales Manager

- Increased [redacted] to \$100 million.
- Doubled sales per representative from \$5 million to \$10 million.
- Suggested new products that increased earnings by 23%.

Why we DID NOT use other references

Objective [Click here and type objective]

Experience 1990-1994 Arbor Shoe Southridge, SC
National Sales Manager

- Increased sales from \$50 million to \$100 million.
- Doubled sales per representative from \$5 million to \$10 million.
- Suggested new products that increased earnings by 23%.



Types of Document Annotation (3)

- Adding knowledge to documents (ctd.)
 - Document enrichment: helping connecting the document to the rest of the knowledge
 - Associating Services
 - Magpie (Domingue *et al.* 2004)
 - Connected to other documents
 - e.g. Automatic generation of hyperlinks
 - COHSE (Goble *et al.* 2001)
 - Magpie (Dzbor *et al.* 2003)
 - AktiveDoc (Lanfranchi *et al.* 2005)

Application Areas for Annotation

■ Annotation Services

- Automatic integration of dispersed information
- Better Indexing and retrieval

■ Knowledge Management

□ Organization's repositories as mini Webs

- Aerospace Boeing, Rolls Royce
- Automotive Fiat
- Biomedicine GlaxoSmithKline, Merck, NPSA
- Services Royal Mail
- KM Quinary (I), Ontoprise(D)
- Other Italian Parliament

CREAM

Siegfried Handschuh and Steffen Staab, "Annotation of the Shallow and the Deep Web, "<http://citeseer.ist.psu.edu/580187.html>

- Annotation framework for ontology-driven annotation
 - Reference implementation is Ontomat
 - <http://annotation.semanticweb.org/ontomat/index.html>
- It supports:
 - Manual annotation of documents
 - Authoring of documents: authors can create annotation while putting together the content of a page
 - Semi-automatic annotation: to reduce the burden of manual annotation
 - Deep annotation: to annotate the Deep Web (documents hidden in databases)
 - When the database owner is cooperatively participating in the Semantic Web.

CREAM: Requirements

- Easy of use and efficiency:
 - Annotation is a difficult task that must be made easy (or easier) for the generic user
 - See next slide
- Ontology based:
 - Ontology provides the interlingua for the Semantic Web (see previous lectures)
- Unique referencing for individuals
 - E.g. “Dieter Fensel” must always be tagged with a unique id in the Knowledge Base, otherwise it won’t be possible to retrieve all the knowledge about him when querying
 - Cream must provide help in retrieving/identifying proper ids

Easy of use?

- The following statement is not exactly easy to write/understand

```
<rdf:Description rdf:ID="CIT1111">  
  <rdf:type rdf:resource="http://www.mydomain.org/uni-ns#course"/>  
  <uni:courseName>Discrete Maths</uni:courseName>  
  <uni:isTaughtBy rdf:resource="#949318"/>  
</rdf:Description>
```

```
<rdf:Description rdf:ID="949318">  
  <rdf:type rdf:resource="http://www.mydomain.org/uni-  
ns#lecturer"/>  
  <uni:name>David Billington</uni:name>  
  <uni:title>Associate Professor</uni:title>  
</rdf:Description>
```

- Parallel with HTML (nobody writes HTML nowadays)
- Need of specialised editors
 - Dreamweaver-like

CREAM: Requirements (2)

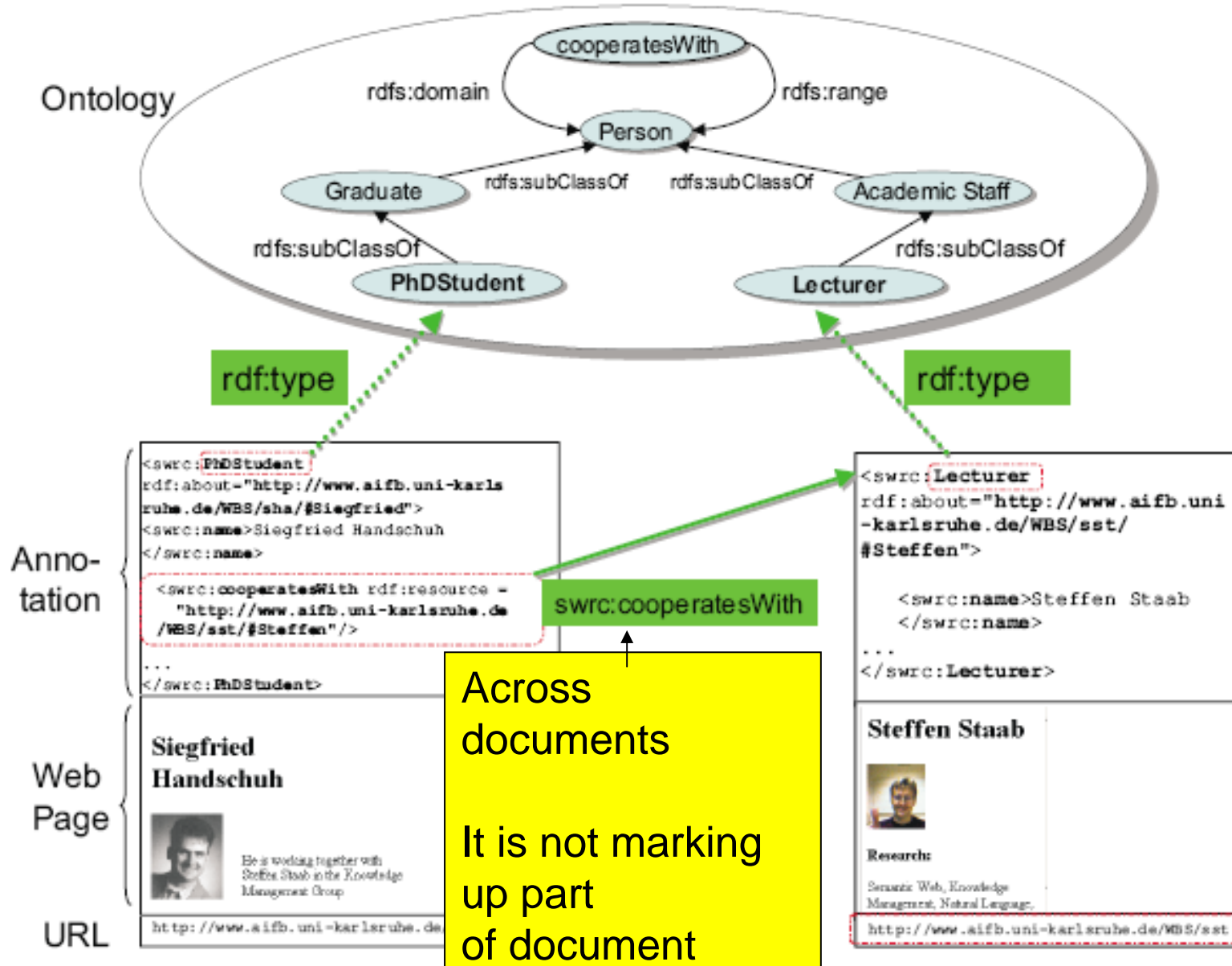
■ Reuse

- Ability to reuse already annotated documents
 - Obvious?

■ Knowledge as a layer that spans across documents

- As hyperlinks connects documents, so knowledge does
 - Connecting knowledge about one individual (using its id)
 - Connecting multiple document content
 - Annotation carries information not in documents!!!
 - See next slides

Knowledge across documents



CREAM: Requirements (3)

■ Maintainability

- Annotation needs maintenance when documents are changed
 - Risk is that document is changed and annotation is not!
- Simplifying maintenance improves quality of SW

■ Multiple ontology

- Possible to annotate using different ontologies
- Supporting different uses of documents

CREAM: Annotation

- The interface enables editing and annotation of documents
- Types of annotations
 - Concepts
 - Properties
 - Relations
- Same objects we have seen in RDF
- HTML Editor with possibility to annotate (in DAML-OIL)
- SW: Annotation:
 - Selection of text and drag and drop is the way in which annotation is performed
 - Easy of use reminds HTML annotation tools
 - E.g. dreamweaver

Ontomat

File Edit View Tools Window Help

Ontology Browser

- AssociateProfessor
- FullProfessor
- Lecturer
- AdministrativeStaff
- Manager
- TechnicalStaff
- Student
 - Graduate
 - PhDStudent
 - Undergraduate
- Product
- Project
- DevelopmentProject

Siegfried Handschuh

Steffen Staab

Rudi Studer

HTML Browser

URL: <http://www.aifb.uni-karlsruhe.de/WBS/sha>

[Knowledge Management Group, Institute AIFB, Karlsruhe University \(TH\)](#)

Siegfried Handschuh

Email: handschuh@aom.org

Email @ Institute: sha@aifb.uni-karlsruhe.de


Phone: ++49-(0)721-608-6554

Fax: ++49-(0)721-608-6580

Office: Kollegiengebäude am Ehrenhof
(Building 11.40)
Englerstrasse 11
2nd floor
room 261

Address: Institute AIFB
University of Karlsruhe
D-76128 Karlsruhe
[Germany](#)

Projects: OntoAgent, PADLR



Ready

1248.0k free

HTML Source Annotation

State: Loaded

Type: text/html

editaole

Text is selected and dropped into a concept in the ontology

Ontology panel

Document panel

Navigation Tree:

- Assistent/Professor
- Professoren
- Lehrer
- Adressbuch
- Manager
- Teaching Assistant
- Student
- Alumni
- Professoren
- Lehrbeauftragte
- Projekt
- Projekt
- Development/Project

Profile Summary:

Siegfried Handschuh

Email: hschuh@l3.informatik.uni-berlin.de

Employer: www.l3.informatik.uni-berlin.de

Phone: +49 10721 908-6004

Fax: 030 20199980

Office: Hollingplatzgebäude am Ehrenhof (Building 11-40) Organisationsstr. 11 2nd floor 10489 Berlin

Address: Institute AFB
University of Applied Sciences
10489 Berlin

Project: OntAgent **PRDLR**

Attribute	Value
age	
email	hschuh@l3.informatik.uni-berlin.de
fax	+49 30 201 908-6004
first_name	Siegfried
last_name	Handschuh
middle_name	
name	Siegfried Handschuh
phone	+49 30 721 908-6004
photo	

Bottom Navigation:

- Startseite
- Navigation
- PRDLR
- OntAgent
- Startseite
- Navigation

Yellow Box 1: The relations and properties of concept are shown

Yellow Box 2: Values are dragged and dropped into proper places

Document generation

- Given the content of the KB, document generation can be helped
 - Canned text is associated to the ontology parts
 - E.g. the relation “<rdf:Property rdf:ID=“collabotares-with”>” has a lexicalization associated, e.g. the string “collaborates with”
 - Dragging a concept/property/relation and dropping it into the text generates automatically
 - Text (using the canned text)
 - The annotation

Internet Explorer window showing a user profile page for Siegfried Handschuh. The browser address bar shows the URL: <http://www.wifi.uni-leipzig.de/PRO/line>.

Siegfried Handschuh

Email: hschuh@uni-leipzig.de
 Phone: +49 (0)721-606-6554
 Fax: +49 (0)721-606-6550
 Office: Kollegiengebäude am Charnock (Building 11-40) Erdgeschoss 11 2nd floor room 201
 Address: Institute AFB, 1st floor of Kollegiengebäude am Charnock, 04109 Leipzig, Germany.

His supervisor is: [Prof. Dr. habil. Gerd Plag](#)

His subordinates are: [Prof. Dr. habil. Gerd Plag](#)

His subordinates are: [Prof. Dr. habil. Gerd Plag](#)

HTML | [Glossar](#) | [Anmeldung](#)

95% Loaded
 Top features
 1541 Bytes

On the left, a navigation tree shows the hierarchy:

- Person
 - AcademicStaff
 - Full Professor
 - Assistant Professor
 - Associate Professor
 - Full Professor
 - Lehrst. (Lehrstuhl)
 - Administrative Staff
 - Manager
 - Technician

 Red arrows and circles indicate the path taken to reach the current page:

- ① Click on "Person" in the navigation tree.
- ② Click on "AcademicStaff" in the navigation tree.
- ③ Click on "Full Professor" in the navigation tree.

Document generation (2)

- It is important to provide tools that help in generating content
 - Annotators will generate annotation (that costs time) if they will have an advantage
 - In terms of better retrieval
 - In terms of content generation help

Annotation of Deep Web

- Deep Web has expected proportion 500/1 with respect to the Shallow Web
 - Search engines do not annotate deep web
 - If a document is there, it cannot be retrieved
 - SW must be able to annotate DW
- Problems in annotating DW
 - Documents are (or can be) generated automatically
 - E.g. responses to queries in eCommerce generates virtual documents using DB content (product, price, etc.)
 - Cannot be annotated singularly
 - They do not exist"

Annotation of DW

- It is necessary to generate rules to annotate the DB schema rather than the individual documents
 - Reverse engineering the DB
 - If DB owner does not collaborate
 - Correlate the DB schema to the ontology if the schema is known
 - Typically when the DB owner cooperates.

Annotation of DW (2)

- CREAM considers the collaborative case
 - The DB schema is considered as another ontology
 - Mapping rules are defined among the ontologies
 - Annotation is inserted using the mapping rules when document is displayed

Issues in User Centred Document Annotation

Annotations: *Where* From?

- SW relies on document annotation
 - Current state of art requires manual annotation
- Manual Annotation
 - Very few people will annotate web pages by hand
 - What if they did?
 - Isn't the web based on hype?
 - Do people really need to publish their girlfriend photos?

Manual Annotation (1)

- Expensive/time consuming/difficult
 - Chicken-egg problem
 - If it adds time to page editing, users will not do it unless there is really something for them
 - Usefulness
 - Hype
- Inefficient and never ending
 - Every new document needs to be annotated
- Difficult
 - if two people annotate the same documents have 15-30/100 probabilities to annotate them differently
 - Risk is that the same information is annotated differently
 - Disagreement between annotators means data sparsity
 - Information becomes difficult to retrieve

An Example

- 10 annotators
- Emails about workshop announcements
 - Name, acronym, date of workshop
 - Name, acronym, URL of associated conference (if any)
 - Submission dates.
- 15% inter-annotator disagreement
 - Especially on name of conference/workshop



- concept
 - Workshop
 - WorkshopName
 - WorkshopAcronym
 - WorkshopDate
 - WorkshopHomepage
 - WorkshopLocation
 - WorkshopPaperSubmissionDate
 - WorkshopNotificationOfAcceptanceDate
 - WorkshopCameraReadyCopyDate
 - Conference
 - ConferenceName
 - ConferenceAcronym
 - ConferenceHomepage
- relation

```

*** Workshop: ***
*** Intelligent Data Analysis in Medicine and Pharmacology ***
*** (IDAMAP 99) ***
*****
Saturday, November 6, 1999
Washington, DC, USA
during the
AMIA 1999 Annual Symposium
November 6-10, 1999 in Washington, DC, USA
(homepage of IDAMAP 99
http://www.ifs.tuwien.ac.at/~silvia/idamap99/
(homepage of AMIA 1999
http://www.amia.org/meetings/f99/call/cover.htm
)
-----
Important dates
* Submission deadline: July 26, 1999
* Notification to authors: September 6, 1999
* Camera-ready paper: October 11, 1999
* Conference: November 6-10, 1999
* Workshop: Saturday, November 6, 1999
-----
GENERAL INFORMATION:
-----
IDAMAP-99 is a Workshop at the AMIA 1999 Annual Symposium - November 6-10, 1999 - Washington, DC prior
to the start of the main AMIA conference.
Gathering in an informal setting, workshop participants will have the opportunity to meet and discuss
selected technical topics in an atmosphere, which fosters the active exchange of ideas among
researchers and practitioners. To encourage interaction and a broad exchange of ideas, the workshop
will be kept small, preferably under 30 active participants, although registered AMIA 99 Fall Symposium
members are welcome to attend. The workshop is intended to be a genuinely interactive event and not a
mini-conference, thus ample time will be allotted for general discussion. The workshop will last a
half-day.
This is the fourth workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP). The
former IDAMAP Workshops were held in Budapest in 1996, in Nagoya in 1997, and in Brighton in 1998.

```

Why not including Annual/Fall symposium?

AMIA 1999 Annual Symposium

Is this the name or the acronym?

Missing workshop location!

Washington, DC

AMIA 99 Fall Symposium

ontology

Problems in the example

- The previous example contains
 - Three doubtful cases (conference name/acronym)
 - One mistake
 - It was annotated by two people and a third one checked their annotations

Problems with Manual Annotation (2)

- Tedious & Tiring
 - Error prone
- Legacy with the past
 - Ontologies are living objects, new version produced
 - Which version of the ontology is used for annotation?
- Dispersed information
 - Annotation largely unfeasible for large diverse repositories
 - E.g. a Web site (Department of CS of the University of Southampton: 1,600 pages)
 - How many relevant ontologies are there for that department?

Problems with Manual Annotation (3)

- How many annotation schemas?
 - The Semantic Web is expected to be composed of
 - [Many] small ontological components [*Hendler 2001*] will be created, mainly related to different domain and applications
 - University of Sheffield web site:
 - What ontology for annotation?
 - Universities/Education, Research life, Scientific Papers,
 - Sport, computer network organization....
 - You name what...

Annotation for use...

- If annotation is to be chosen by author/owner
 - Selection of Annotation Schema may reflect world model of the creator, not of the user
 - E.g. education is the main goal of the university, so the central Uni will probably choose an ontology on Education
 - Most of my time is actually devoted to research
 - Most of my colleagues look for scientific information on our web site
 - To us, Uni's annotation would be largely unuseful
 - Question:
 - Who (and how!) is going to introduce the annotation for us?
 - Where is the annotation to be inserted?

Where to Insert Annotation

- In CREAM annotation becomes part of the document
 - Document is modified
- If a document is annotated by a third party
 - Annotation cannot be inserted in document
 - No permission
 - It must be inserted in a database
 - As current search engine indexes are
 - Used for retrieving/using the page
 - Effect on Semantic Web
 - Annotation may become proprietary
 - As search engine indexes are
 - As any editing done by people (?)

Manual Annotation (2)

- Trusting Manual Annotation?
 - User (in)competence can limit the usefulness of the annotation
 - Spam/Devious
 - Google
 - Does not even use HTML meta Tags! (quality)
 - Avoids using user-defined words only to index (spamming)
 - If we use owner's annotation we are back to the pre-Google world
 - Can be not updated when document is modified
 - If annotation is kept separate from the document
 - e.g. in a database

Automating Annotation for the Semantic Web

Annotation Engines

- Manual document annotation is still largely expected to be the main SW vehicle creation
 - Especially for trusted environment (e.g. within a company) this is expected to provide high quality material
- Automatic annotation is a vision
 - To help manual annotation OR
 - To replace human annotators
- Producing automatic annotation services
 - For a specific ontological component/application
 - Constantly re-indexing documents

Advantages

- Effects:
 - No legacy with the past
 - Annotation with the latest version of the ontology always available
 - Multiple annotation schemas for a single document possible
 - Initial (user) annotation loses importance
 - It is not the only one available, so I can still get information even if the initially associated ontology is irrelevant to me!
 - Simplifies maintenance
 - Page changed but not re-annotated would never happen anymore
 - Like today's search engines cope with disappearing links
 - No annotation in the document
 - The engine would have its database of annotations
 - They are not the page owners, cannot modify your documents!
 - As today's indexes are not stored in the documents

Automatic Document Annotation

- **Ontology based annotation**
 - **User centred**
 - MnM (Vargas-Vera *et al.* 2002)
 - S-Cream (Handschuh *et al.* 2002)
 - Melita (Ciravegna *et al.* 2002)
 - AktiveDoc (Lanfranchi *et al.* 2005)
 - **Unsupervised Domain independent**
 - SemTag and Seeker (Dill *et al.* 2003)
 - Pankov (Cimiano *et al.* 2004)
 - **Unsupervised Domain dependent**
 - Armadillo (Ciravegna *et al.* 2004)
- **Connecting documents to the knowledge space**
 - Magpie (Domingue *et al.* 2004)
 - AktiveDoc (Lanfranchi *et al.* 2005)

Dimensions of Automatic Annotation

- User centred versus unsupervised
- Large Scale (millions) versus small scale (dozens of documents)
- Application-specific versus generic
- Shallow versus Deep
 - **Shallow: Named entity recognition**
 - With recognition of entities across documents
 - Who is Michael Jackson?
 - **Deep: Complex fact capturing**
 - Also across documents
 - E.g. as in Cream
- Supervised versus unsupervised
- Media:
 - Single media (e.g. text only)
 - Multi-media (evidence in each media is considered, evidence is fused in a Boolean way)
 - Cross-media (evidence is searched across media and compared across them)

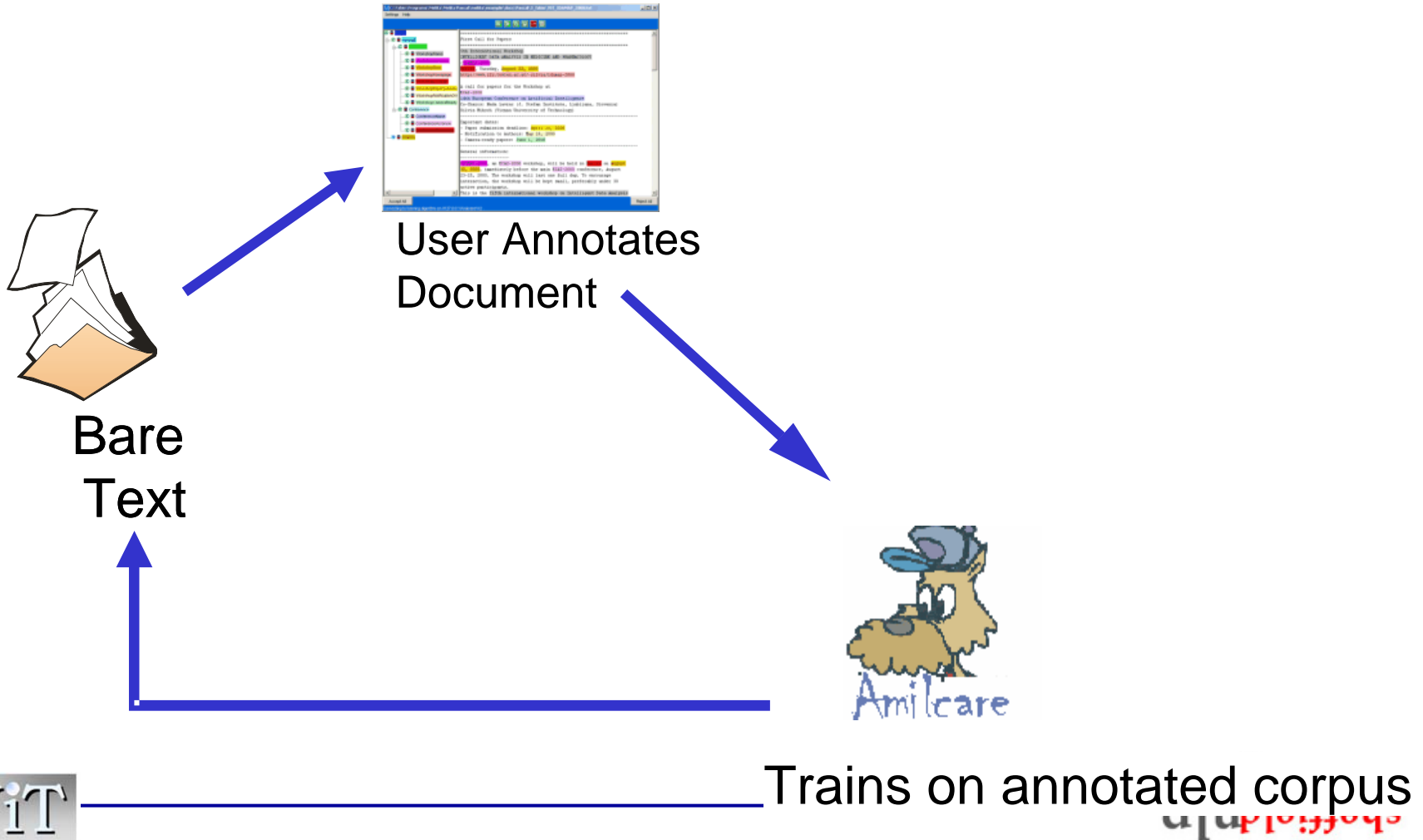
Supporting User-Centred Annotation

- CREAM, MnM and Melita provide semi-automatic annotation
 - Using Machine Learning based IE (Amilcare)
 - To simplify the burden of document annotation
- For trusted environments (e.g. KM)
- Users:
 - Annotates document samples
- IE System:
 - Trains while users annotate
 - Generalizes over seen cases
 - Provides preliminary annotation for new documents
- Advantages
 - Annotates trivial or previously seen cases
 - Focuses slow/expensive user activity on unseen cases
 - Validating extracted information is simpler & less error prone
 - Machine Learning based: it learns how to improve capabilities

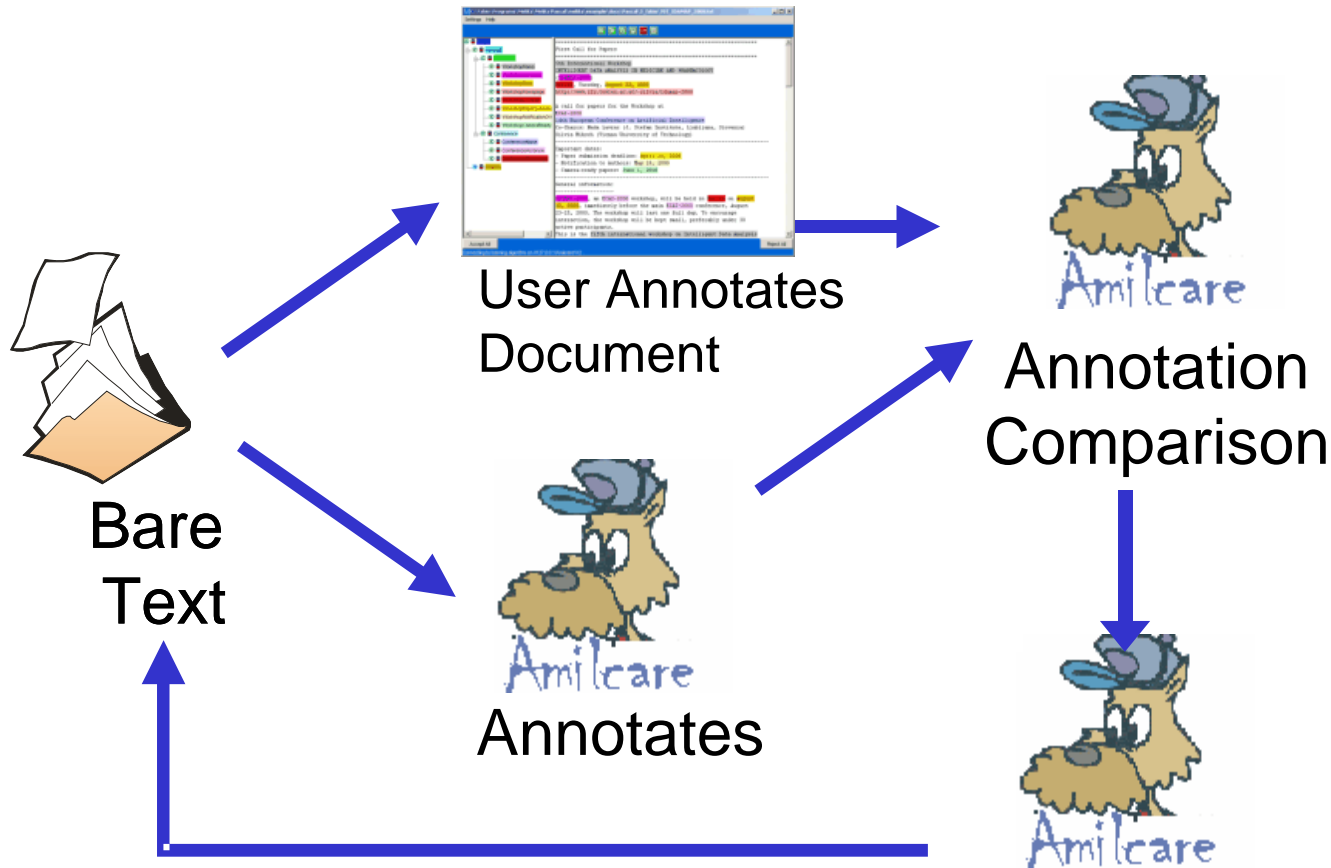
- Ontology-based document annotation assisted by adaptive IE

The screenshot displays the Melita software interface. The title bar shows the file path: `M C:\University\Melita\vr2\Demo\Demo1\Data\docT12.txt`. The menu bar includes `Settings` and `Help`. Below the menu bar is a toolbar with navigation and control icons: a left arrow, a right arrow, a copy icon, a save icon, and a red `OFF` button. The main window is divided into two panes. The left pane, titled `Things`, shows an ontology tree with categories: `Concept` (containing `Speaker` and `Location`), `Time` (containing `Stime` and `Etime`), and `Relation` (containing `At time` and `In location`). The right pane displays document metadata and text. Metadata includes: `Type: cmu.cs.scs`, `Topic: Undergrad Research Presentations`, `Dates: 4-May-92`, `Time: 3:30 - 5:00` (with `3:30` and `5:00` highlighted in yellow and green respectively), and `PostedBy: mjs+ on 29-Apr-92 at 00:39 from G.GP.CS.CMU.EDU (Mark Stehl)`. The `Abstract:` section contains the following text: `The other Independent Study Projects to be presented from 3:30 to 5 on May 4 in Wean 5403 are as follows`, followed by a list of names and topics: `Bill Adams Genie Error Interface Design`, `Gerard Decatrel Virtual Reality: Object Reconstruction`, `Andrew Dent Implementing Biological Imaging Algorithms`, and `Melissa Goldman Ezmail + Dectalk = Eztalk`. At the bottom of the window, there are `Accept All` and `Reject All` buttons, and a `Learning ...` label.

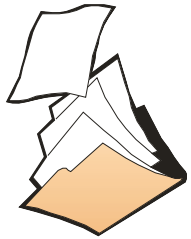
Active Training in Melita (1)



Active Training in Melita (2)



Active Annotation in Melita



Bare
Text



Annotates



User
Corrects



Uses
corrections to
retrain



IE System & Annotator Interplay

- IE system annotates docs
 - Melita uses for suggesting
- Suggestions presented
 - According to the certainty
 - According to user profile
 - Reliable suggestions:
 - Presented in full block
 - Saved if not clicked
 - Fairly reliable suggestions
 - Presented as surrounding boxes
 - Removed if not clicked
- Users can customize system behaviour
 - Intrusivity minimization

The screenshot displays the 'Things' tree on the left and a text document on the right. The 'Things' tree is structured as follows:

- Things
 - Concept
 - Speaker
 - Location
 - Time
 - Stime
 - Time
 - Relation
 - At time
 - In location

The text document on the right contains the following information:

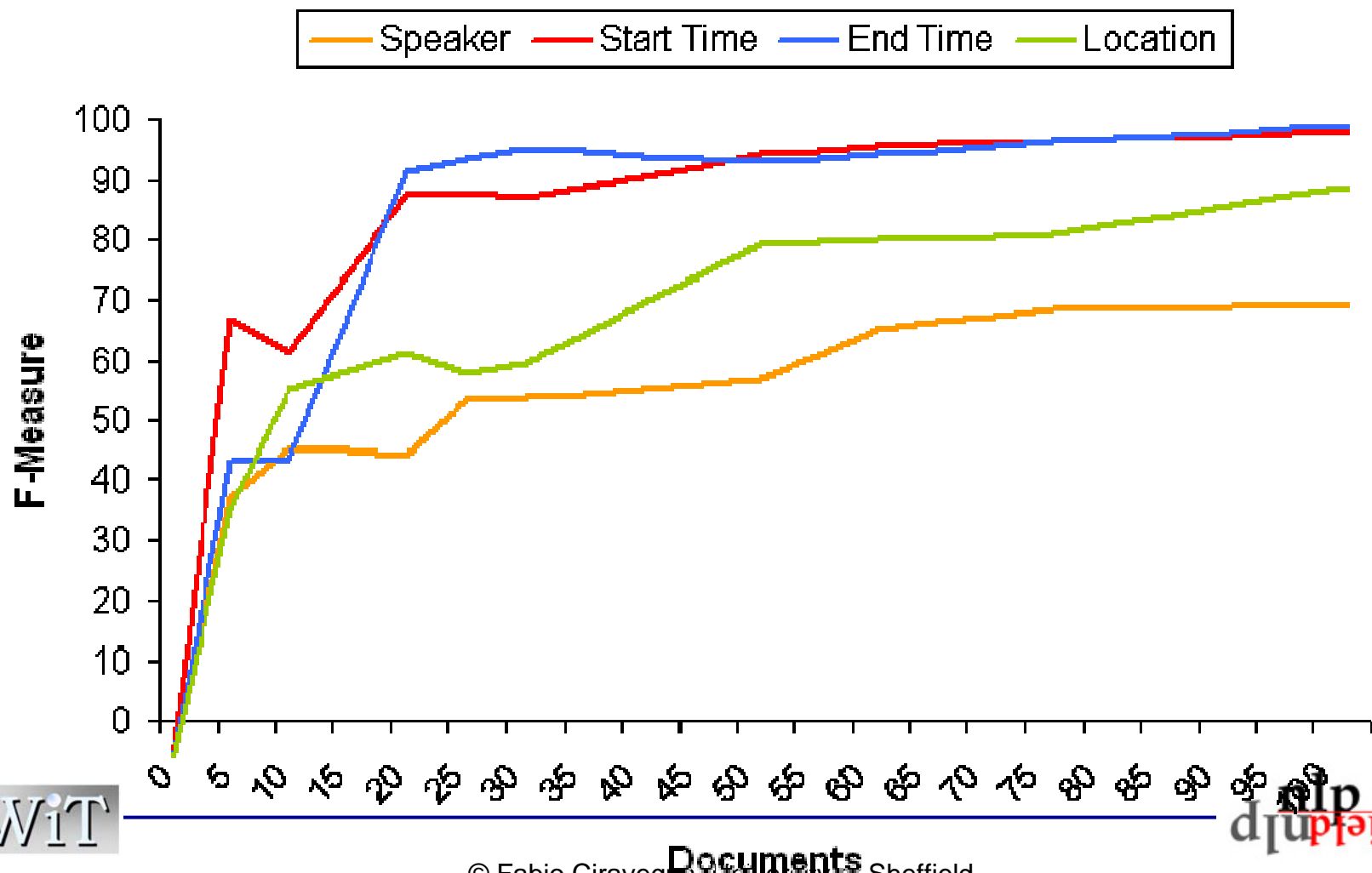
Type: cmu.cs.scs
Topic: Undergrad Research
Dates: 4-May-92
Time: 3:30 - 5:00
PostedBy: mjs+ on 29-Apr-92 e
Abstract:
The other Independent Study F
May 4 in Wean 5403 are as fol
Bill Adams Genie Error
Gerard Decatrel Virtual Real
Andrew Dent Implementing
Melissa Goldman Ezmail + Dec

Dotted arrows indicate the mapping between the 'Things' tree and the text document. For example, 'Speaker' points to 'Bill Adams', 'Location' points to 'Genie Error', 'Stime' points to '3:30', 'Time' points to '5:00', 'At time' points to 'May 4', and 'In location' points to 'Wean 5403'.

Quantitative Support

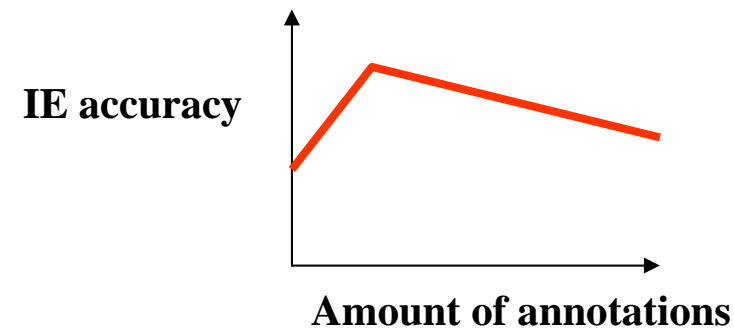
- How quickly does it learn?
- Experiment:
 - Seminar announcements at Carnegie Mellon University
 - Emails to be annotated with
 - Speaker
 - Start time of seminar
 - End time of seminar
 - Location of seminar
 - Note: not as simple as it seems
 - Many people, locations and dates in announcements:
 - Task is spotting the right ones

Quantitative Support in Annotation (how quickly does it learn?)

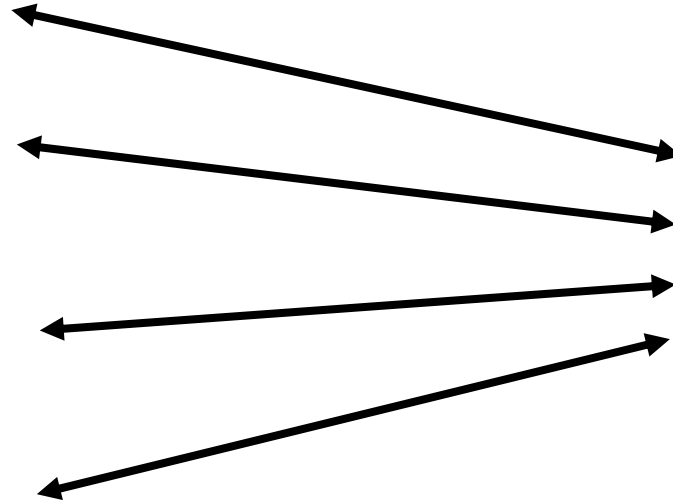


Impact on Annotation

- University of Karlsruhe's experiments (Cream)
 - -80% annotation time
 - +100% interannotator agreement
 - Is this positive?
- Outstanding issue:
 - Impact on annotators of suggestions topping 85% accuracy?



Architecture



Melita Clients

- Invoke IE
- Provide annotated corpora
- Sort documents

IE Server

- Learns from example
- Annotates documents

Application Areas

■ Knowledge Management

- Aerospace Boeing
 - Biomedicine NPSA, Merck, NHS
 - Intelligence MET, SAIC
Lawrence Livermore National Laboratory
 - Law Quinary
- **Solcara:**
- **Next version of KM tool will include Melita & Amilcare**



Amilcare *(Ciravegna 03)*

- Based on $(LP)^2$ algorithm *(Ciravegna 2001)*
- Trains on documents XML annotated
- Integrated with annotation tools:
 - **MnM (Open Univ.), Ontoannotate (Ontoprise, DE)**
 - **Ontomat (Karlsruhe Univ.), SemantiK (Quinary, I)**
 - **Melita (Sheffield Univ.)**
- Limited distribution: released to about 50 sites:
 - **Industrial or Commercial Sites:**
SAIC (Usa), Max Planck Institute (D), Merck (D), Solcara (GB), Lawrence Livermore National Laboratory (Usa), Boeing (Usa), GlaxoSmithKline (Usa), Quinary (I), Ontoprise (D), Mondeca (F), Camera dei Deputati (Italian Parliament) (I)
 - **Academic Sites:**
University College Dublin (IE), CNRS (F), University of Cambridge (UK), University of Trier (D), NCRS Demokritos (Gr), Carnegie Mellon University (Usa), University of Illinois (Usa), University of Texas, Austin (Usa), Open University (UK), Danmarks Tekniske Universitet (Dk), University of Southampton (UK), Arizona State University (Usa), Naval Postgraduate School Monterey (Usa),

Connecting to the Knowledge Space

- Annotating single documents is not enough
- As CREAM shows
 - There are many cases where it is necessary to connect knowledge in different documents
 - Using unique IDs
 - But also to refer to already known knowledge (knowledge reuse)
 - Knowledge Reuse
 - From personal knowledge to an organization's knowledge
 - Recovering the context of a document
 - Adding knowledge not present in the document

Magpie

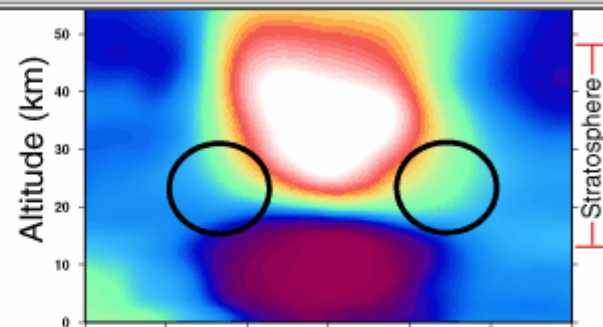
Martin Dzbor, Enrico Motta, John Domingue, Marc Eisenstadt

“MagPie, A tool for the SW” <http://kmi.open.ac.uk/projects/magpie/main.html>

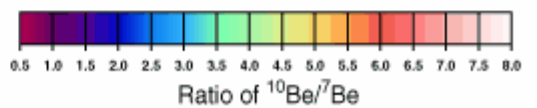
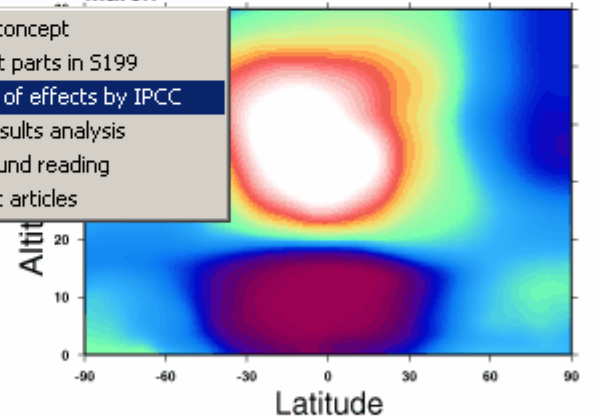
- Magpie enables opening up the knowledge space of the document by connecting the contained knowledge to the outside world
 - E.g. contained concepts are automatically hyperlinked to their definition
 - Individuals are linked to their id in the KB
 - Named entity recognition
 - Services can be associated to concepts in the ontology
 - Services are used to display further information

collision of high-energy particles from space with nitrogen atoms in the atmosphere. Most tracer production occurs between about 30° 70° latitude in both hemispheres of the lower stratosphere, as indicated by the circled regions on the figure. These tracers, which are borne on aerosol particles, are removed from the stratosphere by radioactive decay. While beryllium-7 decays relatively quickly, with a half-life of 53 days, ¹⁰Be's decay rate is negligible. The only sink for ¹⁰Be occurs after it enters the troposphere, where the radionuclides are efficiently removed by precipitation. Therefore, if we look at the ratio of ¹⁰Be/⁷Be as air moves from the midlatitude production region to other parts of the stratosphere, the ratio will generally increase, as ⁷Be decays. Thus, the ¹⁰Be/⁷Be acts as a "clock" of air mass age.

The figure shows the ¹⁰Be/⁷Be ratio calculated in the GISS general circulation model (GCM) during January and March. In the tropical stratosphere, air rises from the troposphere and continues to ascend, but exchange with higher latitudes is inhibited. The ¹⁰Be/⁷Be ratio is very high (white region) since slow penetration of air from the midlatitude production region allows much of the ⁷Be to decay. During the early northern hemisphere spring, air from the lower tropical stratosphere moves to higher latitudes relatively quickly. The result is the green blob of relatively high ¹⁰Be/⁷Be air at



March



¹⁰Be/⁷Be ratio calculated in the GISS general circulation model during January and March. Circled areas indicate maximum

- Explain concept
- Relevant parts in 5199
- Analysis of effects by IPCC
- CPDN results analysis
- Background reading
- Scientific articles

Annotation in AktiveDoc

- Document Editor/Browser for SW
- It covers the three levels of annotation
 - Ontology-based
 - Braindump (comments a la Word)
 - Expansion of knowledge space (a la Magpie using large scale IE)
- It provides suggestions for content taking into account the context being written
 - Extracting content a la Melita
 - Searching the SW (e.g. knowledge bases)
- It provides privacy and security of annotation
 - Does not modify the document
 - Annotation in a database
- Services associated to annotated concepts

Ontology

- information_bearing
- attending_a_conferen
- generic_agent
- r_and_d_institute
- charitable_organiza
- multimedia_designe
- attending_an_event
- organization_unit
- employee
- conferring_an_awar
- serial_publication
- learning_centred_o
- educational_organiz
- generalised_means
- event_involving_mc
- book
- operating_system
- higher_educational
- thesis_reference
- event_involving_pri
- city
- article_reference
- industrial_organizat

Arial 1 (8 pt) Heading 1 **B** *I* U **S** x_2 x^2

Use Amilcare Reject Suggestion

Search with Google Search with Armadillo Search in local domain

Insert a comment

We also plan to have a lively panel discussion around these topics and Christopher is organising participation and contacting interested AKTors. Panclists will be finalised next week when we all meet for the joint demo session, on Tuesday. So, far the following AKTors have expressed interest to attend both or one of the sessions: **Fabio Ciravegna** (shef), Craig McKenzie (abdn) Dave Robertson (edin) **Enrico Motta** (ou) Vita Lanfranchi (shef) Chen-Burger (aoton) Christopher Brewster (shef) Sam Chapman (shef) Harith Alani (soton) Mischa Tuffield (soton)

We are aware that some people will start their decent to Southampton over the weekend, so we are keen to finalise an

Path: [body](#) > [img](#)

Save in directory SAVE CANCEL SELF REF. REF. BY DATE

Full Title: Dr
 Email: F.Ciravegna@dcs.shef.ac.uk
 Telephone number: + 4401142221940
 Home Page:
<http://www.dcs.shef.ac.uk/~fabio/>







Large Scale Annotation

- One step further
 - Towards large scale annotation
 - Many document sources (sites)
 - Variety
 - Consistency
 - Dispersed information (no self contained documents)
 - Information integration needs
 - Human-centred annotation largely unfeasible
 - OR...?
- Proposal: automatic annotation services

Dimensions of approach classification

- Task:
 - Shallow versus Deep
 - Named entities versus event extraction
- Ontologies
 - Generic ontologies versus application specific
 - Scale (often directly proportional to genericity)
 - Large ontologies (e.g. TAP: 10,000s of concepts)
 - SemTag (Dill *et al.* 2003)
 - Versus application specific ontologies (100s of concepts)
 - Armadillo (Ciravegna *et al.* 2004)
- Requirements:
 - To enable automatic processing
 - Requirement: High accuracy (as in databases)
 - To enable human centred searching
 - Requirement: medium accuracy (as in web searches)

Armadillo

Used in the Hands on Session!



- System for Large Scale Annotation
 - Capturing events
- Composable architecture
- Annotation as Harvesting
 - Searching, Classifying, Extracting, Integrating, Visualizing
- Ontology based
 - Ontology defines application domain (dozens to hundreds of concepts)
- Uses an RDF triple store to store extracted facts
- Supports geographically distributed architectures

Annotation as Harvesting

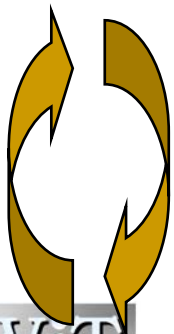
- Harvesting defined as:
 - Task of identifying instances for objects in a given ontology
 - Both entities and relations
- Harvesting modules
 - Defined according to objects they work on
 - Formally defined in terms of the task(s) they perform
 - E.g. classification, extraction, integration, visualization...
- Information Food Chain metaphor (*Etzioni96*)
 - Search engines/classifiers as herbivores
 - Armadillo uses existing search systems (for Web or company repositories)
 - Information agents as carnivores
 - Information Extraction
 - Information Integration

Extraction

- To model Deep Web
 - Models the data base schema to the ontology
 - As in Cream
 - Rules to wrap existing regular web sites
 - E,g, automatically generated by a database
 - What you will do in the hands on session using regular expressions!
- To extract from generic web pages
 - Semi-supervised approach
 - Next slide

Large Scale Extraction Strategy

- Redundancy to bootstrap unsupervised learning
 - **Starting point:**
 - **Seed examples provided via**
 - **user-defined lexica**
 - **easy to model/mine sources (wrappers)**
 - **Armadillo**
 - **Searches mentions in corpus**
 - **Multiple strategies to combine evidence**
 - Is this really its instance?
 - **Cycle:**
 - **Seed examples used to bootstrap learning**
 - For progressively more complex cases
 - From lists and tables to free text
 - **Produces more examples**
 - Multiple strategies to combine evidence



Not used in the Hands on Session!

Information Integration

- Facts from different sources need to be integrated
 - To connect information/knowledge
 - To solve discrepancies and ambiguities
- Steps
 - Unique instance identification (for entities)
 - Record linkage (for events)
- Information Integration strategies
 - Generic
 - Distance metrics
 - Used in the HandsOn!
 - Using Web bias
 - Application specific
 - Rules

Gourm-adillo



Armadillo : Est Est Est - Edinburgh

[Natural Language Processing Group](#),
[Department of Computer Science](#),
[University of Sheffield](#).

Regent Court, 211 Portobello Street, Sheffield, S1 4DP,
UNITED KINGDOM

Tel: +44(0)114-2228000 Fax: +44(0)114-22-21810

sam@dcs.shef.ac.uk

Links

[Sam's HomePage](#)

[Armadillo](#)

This page details the Information Extracted regarding the Restaurants, (Est Est Est - Edinburgh) and related details for it. All details on this page have been automatically extracted from existing web resources by the Armadillo Information Retrieval and Information Extraction tool Then integrated to create this new web portal. All content on this page is sourced from remote web resources and where applicable the URL's of the resources are indicated. The owner of this site does not present these results as fact but as the results of automatic extraction therefore we hold no liability for any errors or omissions

[Back to main Restaurants Index](#)

Est Est Est - Edinburgh

DESCRIPTION: As soon as you walk through the door of an Est Est Est outlet you should instantly feel welcome. It is a vibrant modern Italian restaurant offering a sensibly priced, up to date menu in stylish surroundings with a warm and friendly atmosphere, where a simple one course meal can cost as little as £10. They try hard to cater to all tastes, some restaurants feature wood-fired pizza ovens, which produce sensational pizzas far larger and tastier than most others you'll find on the high street. The menu offers a great range of chicken and steak dishes, and seafood too - all added to each week by a range of blackboard specials created by the chef. Of course, being an Italian restaurant, Est Est Est serves an excellent selection of pasta, piled hot and steaming into large bowls - perfect for lunch or early evening with a crisp side salad. Children are particularly well catered for at Est Est Est. The menu is varied and even encourages participation, for instance they can top their own pizzas with their own choice of toppings; and all children's main course prices include ice cream and a soft drink. Further Information [Est Est Est is a ...](#)

Est Est Est - Edinburgh - Maps



ArTmadillo

- Mines the web to retrieve information on painters and their works

Armadillo : Michelangelo Merisi da Caravaggio

This page details the Information Extracted regarding the artist, Michelangelo Merisi da Caravaggio and artworks produced by them. All details on this page have been automatically extracted from existing web resources by the Armadillo Information Retrieval and Information Extraction tool. All content on this page is sourced from remote web resources and where applicable the URL's of the resources are indicated.

[Back to sub artist Index](#)
[Back to main artist Index](#)
[Back to main artwork Index](#)

Full Name:	Michelangelo Merisi da Caravaggio
First Name:	Michelangelo
Surname:	Caravaggio

Image of Artist:




Image extracted from http://www.artrenewal.org/images/artists/c/Caravaggio_Michelangelo_Merisi_da/biopic.jpg
This Image may be unavailable as it links to the remote web resource which may require the correct cookies of referring page to access the image.

ArTmadillo

Beheading of Saint John the Baptist by Michelangelo Merisi da Caravaggio - Netscape

Beheading of Saint John the Baptist by Mic...

Image of Artwork:




Image extracted from http://www.artrenewal.org/images/artists/c/Caravaggio_Michelangelo_Merisi_da/large/Beheading_of_Saint_John_the_Baptist_WGA. This Image may be unavailable as it links to the remote web resource which may require the correct cookies of referring page to access the image.

Production Date:	1608
Methods used to obtain Data:	IR
Armadillo Validity Rating:	0.9

W

shettie

Artists domain Evaluation

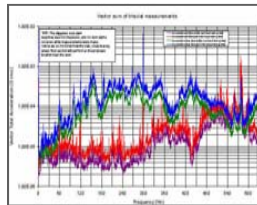
Artist	Method	Precision	Recall	F-Measure
Caravaggio	II	100.0%	61%	75.8%
	IE	100.0%	98.8%	99.4%
Cezanne	II	100.0%	27.1%	42.7%
	IE	91.0%	42.6%	58.0%
Manet	II	100.0%	29.7%	45.8%
	IE	100.0%	40.6%	57.8%
Monet	II	100.0%	14.6%	25.5%
	IE	86.3%	48.5%	62.1%
Raphael	II	100.0%	59.9%	74.9%
	IE	96.5%	86.4%	91.2%
Renoir	II	94.7%	40.0%	56.2%
	IE	96.4%	60.0%	74.0%

Future of Annotation

- What's next?
 - Text only?
 - Multimedia?
 - Cross-media?
- Industrial use?
 - Is there any industrial use of annotations yet?
- X-Media as an example of project
 - Integrated Project
 - Coordinated by University of Sheffield
 - >€10M funding
 - Currently under final negotiation

X-MEDIA

Knowledge Sharing and Reuse across Media



University of Ljubljana



Rolls-Royce



U Freiburg



The project

Our Vision:

A new Approach to KM across Media in Complex Distributed Environments

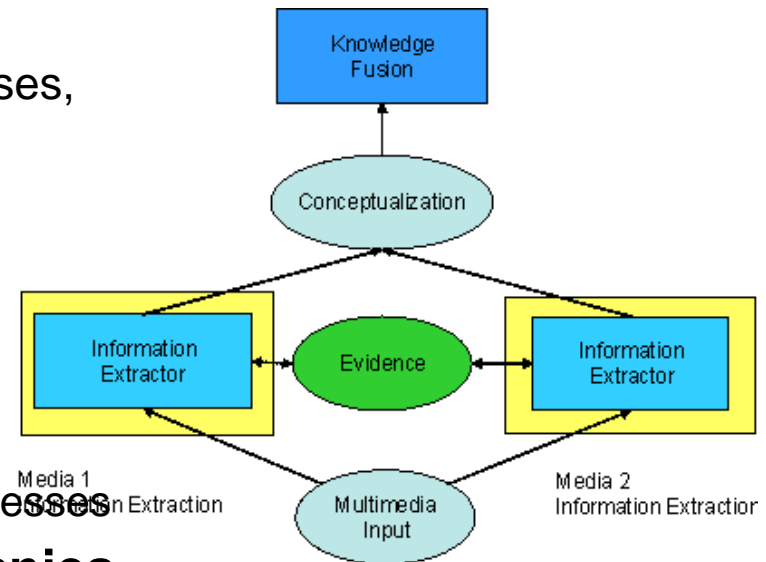
■ Large Scale Acquisition, Sharing and Reuse of Knowledge

- Distributed in images, documents and data
- Distributed in different repositories (data bases, knowledge bases, etc.)
- Inaccessible for current systems because Knowledge implicit across media.

■ X-Media Technology

- To enable Know How retention / exploitation
- To maintain and improve competitiveness
- To manage knowledge-intensive complex processes

■ Know-how as main asset for EU companies



Main Innovations

■ Application-level innovation

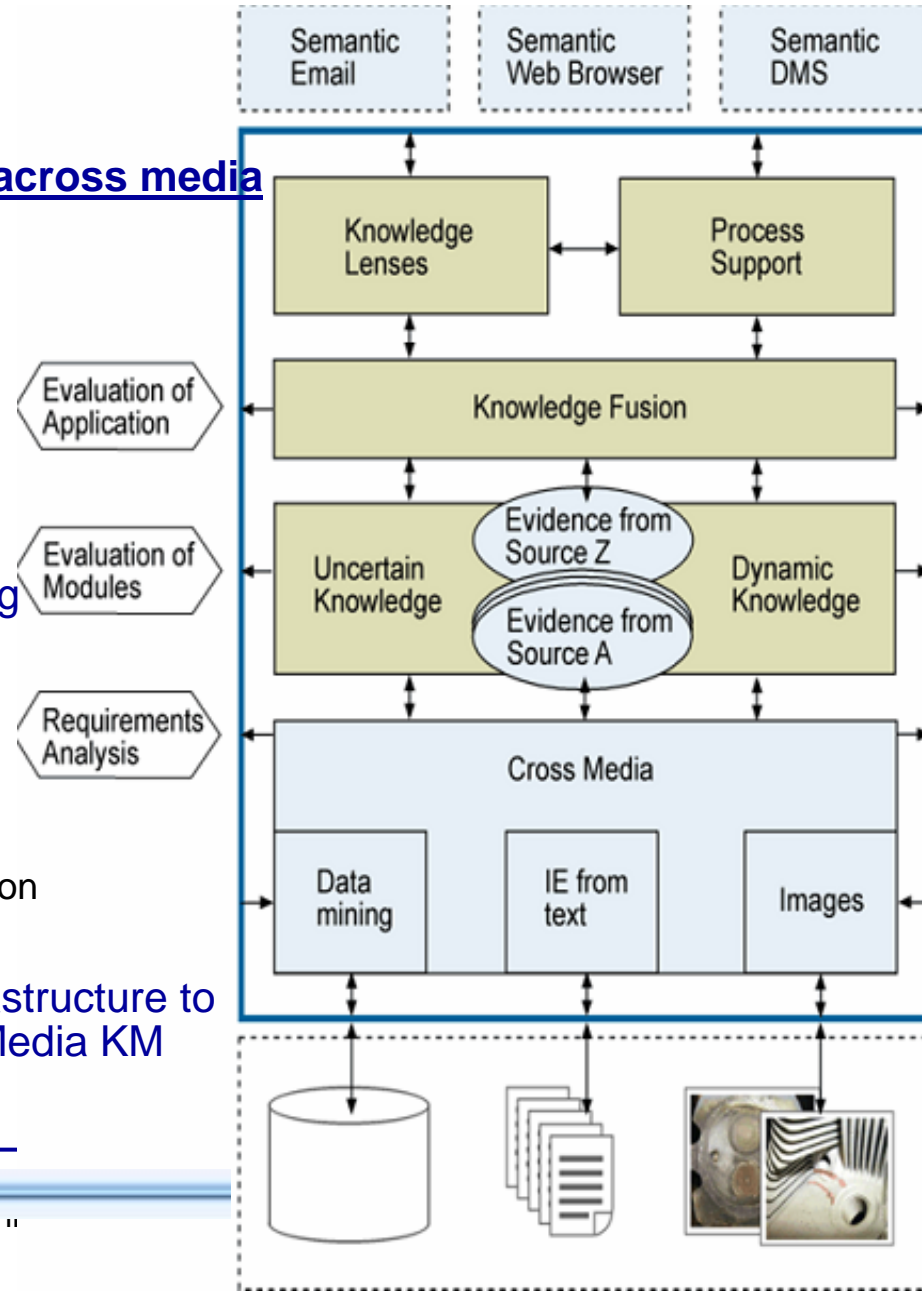
- Enabling new ways of managing knowledge **across media**
 - Currently impossible (size, heterogeneity)
- Exemplar knowledge management systems
- Application for industrial partners:
 - KM for jet engines (RR)
 - Competitor analysis (Fiat).

■ Basic research

- Knowledge Lenses
- Probabilistic Ontologies for rigorous modelling
 - Uncertainty & trust and provenance
 - Dynamic aspects of knowledge
- Approaches to knowledge fusion able to:
 - Merge contradictory knowledge across media
 - Enable users to judge result confidence
 - Algorithms for **cross media** knowledge acquisition

■ System-Level Research

- Methodology, architecture and technical infrastructure to Integrate algorithms/techniques into Cross Media KM systems.



Testbeds

- The multimedia field identified (texts, images and data) fits well an environment in which
 - (1) Sensors and cameras provide basic data to be interpreted
 - (2) Textual documents complement, describe, and help interpret data and images and (3) ontologies describe the domain and the application
- Testbeds
 - Product lifecycle monitoring (Rolls Royce)
 - Competitor analysis (Fiat).

Conclusions

- Document annotation can be performed at different levels
 - **Ontology-based, braindump, document enrichment**
- Annotation unlikely to be performed manually on a large scale except for limited cases (e.g. FoaF)
- Automation can be applied successfully for helping annotating
- We have seen:
 - **User centred automated ontology-based annotation**
 - **For trusted self contained documents (e.g. KM)**
 - **Melita/SCREAM**
 - **Automatic document Enrichment**
 - **Magpie/AktiveDoc**
 - **Unsupervised large scale annotation**
 - **For distributed large scale environments (e.g. the Web)**
 - **Armadillo**

Future Work & Challenges

- Multidisciplinary research for automation
 - NLP has strong role, but complemented with other disciplines
 - SE, ML, II, SWS, HCI
- Annotation
 - Beyond the division between user centred and unsupervised
 - Strong HCI strategies
 - Validation of results across documents
 - How can you validate 2M triples produced by large scale annotation?
- Information extraction models
 - Beyond simple IE models
 - Towards fully fledged adaptive IE systems
 - Maintaining flexibility
- Information Integration
 - Towards complex trainable strategies for integration
- Combination of evidence
 - Of sources
 - Of extractors

Future Work & Challenges (2)

- How modelling uncertainty?
- Knowledge is dynamic. How do you model that?
- HCI
 - Information presentation (document annotation)
 - Intrusivity:
 - **How to avoid annoying users with too many annotations**
 - Trust
 - **Who do users trust?**
 - Tracing preferred sources
 - **Where does the information come from?**
- Scalability
 - Large scale indexing systems
 - Millions of pages (not billions!)

Final thought

The Karen Spark-Jones Final Slide!

These technologies allow easy collection of large amount of information/knowledge

- Are we:
 - Preparing for a better world?
 - Preparing for a world with no secrets/privacy?
 - Big brother
 - Spam
 - Just adding hay to the haystack while searching for a needle?

Thanks to:



www.aktors.org

- Yorick Wilks
- Christopher Brewster
- Sam Chapman
- Ajay Chakravarthy
- Alexiei Dingli
- David Guthrie



www.dot-kom.org

- Neil Ireson
- Jose` Iria
- Barry Norton
- Vita Lanfranchi
- Mark Stevenson

www.3worlds.org

- Vita Lanfranchi
- Daniela Petrelli

Thank You

- Contact Information
 - F.Ciravegna@dcs.shef.ac.uk
 - www.dcs.shef.ac.uk/~fabio
- Web Intelligence
 - <http://nlp.shef.ac.uk/wig/>
- NLP Sheffield
 - <http://nlp.shef.ac.uk/>
- AKT Project
 - www.aktors.org
- Dot.Kom Project
 - www.dot-kom.org
- IPAS project
 - www.3worlds.org



A very Incomplete Bibliography

Information Extraction

- J. Hobbs. The generic information extraction system. In Proceedings of the Fifth Message Understanding Conference (MUC5), pages 1172–1178, 1993.
- F. Ciravegna: Challenges in Information Extraction from Text for Knowledge Management, in S. Staab, (ed), “Human Language Technologies for Knowledge Management”, IEEE Intelligent Systems and Their Applications (Trends and Controversies), Vol. 16, No. 6, pp 88-90, 2001.
- N. Kushmerick, B. Thomas. Adaptive information extraction: Core technologies for information agents (2002). <http://citeseer.nj.nec.com/kushmerick02adaptive.html>
- Califf and Mooney: Relational Learning of Pattern Matching Rules for Information Extraction <http://citeseer.nj.nec.com/6804.html>
- Bikel D., Schwartza R., Weischedel. R. An algorithm that learns what’s in a name. Machine Learning 34, pp.211-231, 1999
- Fabio Ciravegna. Adaptive information extraction from text by rule induction and generalisation. In Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI), 2001. Seattle.
- H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). 2002.

A very Incomplete Bibliography

Information Extraction (ctd.)

- Roman Yangarber, Ralph Grishman, Pasi Tapanainen and Silja Huttunen: "Automatic Acquisition of Domain Knowledge for Information Extraction" In Proc. of COLING 2000, 18th Intern. Conference on Computational Linguistics, Saarbrücken, 2000.
- [Grishman 1997] Ralph Grishman, "Information Extraction: Techniques and Challenges, . In Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, in M.T. Pazienza, (ed.), Springer, 97.
- Douglas E. Appelt, Jerry R. Hobbs, John Bear, David Israel, and Mabry Tyson. 1993. FASTUS: A nite-state processor for information extraction from real-world text. In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, Chambéry, France.
- Fabio Ciravegna, Alberto Lavelli, Nadia Mana, Luca Gilardoni, Silvia Mazza, Massimo Ferraro, Johannes Matiassek, William J. Black, Fabio Rinaldi, and David Mowatt. 1999. FACILE: Classifying texts integrating pattern matching and information extraction. In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, Stockholm, Sweden
- Aaron Douthatt. 1998. The Message Understanding Conference scoring software user's manual. In Proceedings of the Seventh Message Understanding Conference (MUC-7), http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html
- I. Muslea, S. Minton, and C. Knoblock. 1998. Wrapper induction for semistructured webbased information sources. In Proceedings of the Conference on Automated Learning and Discovery (CONALD), 1998.
- S. Soderland. 1999. Learning information extraction rules for semi-structured and free text. Machine Learning, 34(1):233–272
- D. Maynard, K. Bontcheva and H. Cunningham. Towards a semantic extraction of named entities. Recent Advances in Natural Language Processing, Bulgaria, 2003.

A very Incomplete Bibliography(2)

User-centred Annotation

- Vitaveska Lanfranchi, Fabio Ciravegna, Daniela Petrelli: *Semantic Web-based Document: Editing and Browsing in AktiveDoc*, Proceedings of the 2nd European Semantic Web Conference , Heraklion, Greece, May 29-June 1, 2005
- Handschuh, Staab, Ciravegna. S-CREAM - Semi-automatic CREATION of Metadata (2002) <http://citeseer.nj.nec.com/529793.html>
- [Day et al'97] D. Day, J. Aberdeen, L. Hirschman, R. Kozierok, P. Robinson, and M. Vilain. Mixed-Initiative Development of Language Processing Systems. In Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP'97). 1997.
- [Ciravegna'02] F. Ciravegna, A. Dingli, D. Petrelli, Y. Wilks: User-System Cooperation in Document Annotation based on Information Extraction. Knowledge Engineering and Knowledge Management (Ontologies and the Semantic Web), (EKAW02), 2002.
- M. Vargas-Vera, Enrico Motta, J. Domingue, M. Lanzoni, A. Stutt, and F. Ciravegna. MnM: Ontology driven semi-automatic or automatic support for semantic markup. In Proc. of the 13th International Conference on Knowledge Engineering and Knowledge Management, EKAW02. Springer Verlag, 2002.
- C. Goble, S. Bechhofer, L. Carr, D. De Roure, and W. Hall. Conceptual Open Hypermedia = The Semantic Web? In *The Second International Workshop on the Semantic Web*, pages 44–50, Hong Kong, May 2001.

IE and Semantic Web

- Nicholas Kushmerick, Fabio Ciravegna, AnHai Doan, Craig Knoblock, Steffen Staab: Dagstuhl Seminar on Machine Learning for the Semantic Web, 13-18 February 2005, <http://www.smi.ucd.ie/Dagstuhl-MLSW/proceedings/>
- Fabio Ciravegna. Designing adaptive information extraction for the Semantic Web in Amilcare. In S. Handschuh and S. Staab, editors, *Annotation for the Semantic Web*, Frontiers in Artificial Intelligence and Applications. IOS Press, 2003.
- K. Bontcheva, H. Cunningham: Information Extraction as a Semantic Web Technology: Requirements and Promises. Adaptive Text Extraction and Mining workshop, 2003.

A very Incomplete Bibliography(3)

Harvesting

- Fabio Ciravegna, Sam Chapman, Alexiei Dingli, and Yorick Wilks: Learning to Harvest Information for the Semantic Web, Proceedings of the First European Semantic Web Conference, Crete, May 2004
- A. Kiryakov, B. Popov, et al. Semantic Annotation, Indexing, and Retrieval. 2nd International Semantic Web Conference (ISWC2003), <http://www.ontotext.com/publications/index.html#KiryakovEtAl2003>
- S. Dill, N. Eiron, et al: <http://www.tomkinshome.com/papers/2Web/semtag.pdf> . SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation. WWW'03.
- Thomas Leonard and Hugh Glaser. Large scale acquisition and maintenance from the web without source access. In Siegfried Handschuh, Rose Dieng-Kuntz, and Steffen Staab, editors, *Proceedings Workshop 4, Knowledge Markup and Semantic Annotation, K-CAP 2001*, 2001.

Browsing

- Martin Dzbor, John B. Domingue, and Enrico Motta. Magpie - towards a semantic web browser. In Proceedings of the 2nd Intl. Semantic Web Conference, October 2003. Sanibel Island, Florida.

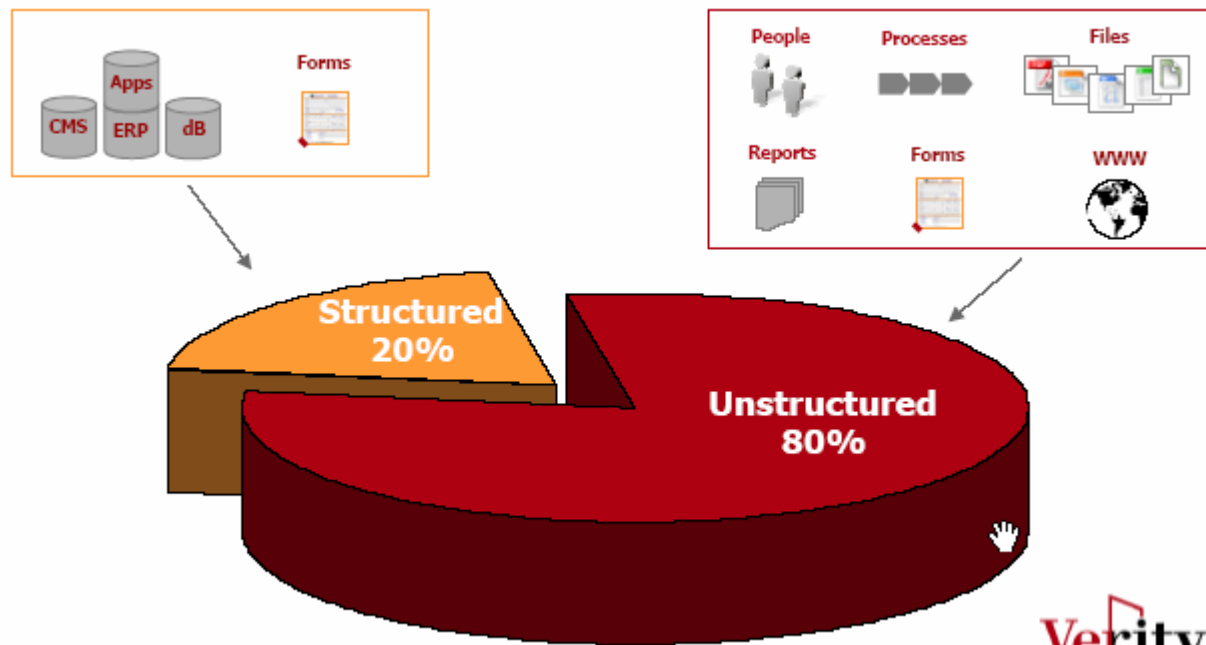
The importance of Managing Unstructured Knowledge

Impact of Limited KM

- International Data Corp. (IDC)
 - Knowledge workers spend from 15% to 35% of their time searching for information.” [KMWorld Volume 13, Issue 3, March 2004].
 - The lack of efficiency costs organizations \$750 billion annually due to wasted time spent by knowledge workers seeking and capturing information necessary for them to do their jobs (A.T. Kearney)
 - Fortune 500 companies lose at least \$31.5 billion a year by failing to share knowledge

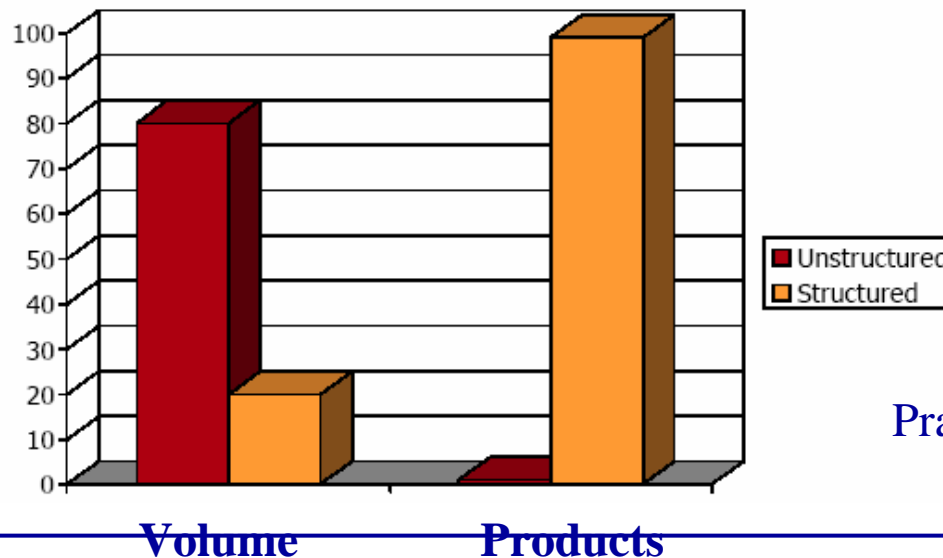
Sources of Knowledge

- 80-85% of a company's knowledge is contained in unstructured form,
 - i.e. expressed in some forms of natural language.



Impact of Knowledge Types

- Content (unstructured information) is much more valuable than structured information (as in databases),
- Availability to companies is generally very limited.
 - Available products tend to provide access to structured rather than unstructured information.



Prabhakar Raghavan (2004)

Expected Industrial Trends

- Strong need for tools to access knowledge through
 - effective and efficient searching,
 - extraction and integration of information
- Businesses spent \$2.7 billion on new systems in 2002,
 - Number to rise to \$4.8 billion in 2007.
- IDC
 - Strong demand for the latest content technologies, including
 - Multimedia and multi-format search and text mining.
 - Content management and retrieval software spending will outpace the overall software market by 2007.
 - Market is estimated at \$6.46 billion market in 2004 and a \$9.72 billion industry by 2006, according to research from IDC.
- Gartner Group
 - 75% of the productivity improvements in corporations should be attributed to introduction of KM practises by 2007
- An important characteristic of unstructured knowledge is its decentralization:
 - Gartner Group: 80% of a company's digital resources are not accessible to the enterprise as a whole
 - they are stored as personal files on individual computing systems, rather than in central repositories. (Computing, 18 November 2002).

Governmental Trends

The market of Knowledge Management (KM) is expanding world-wide

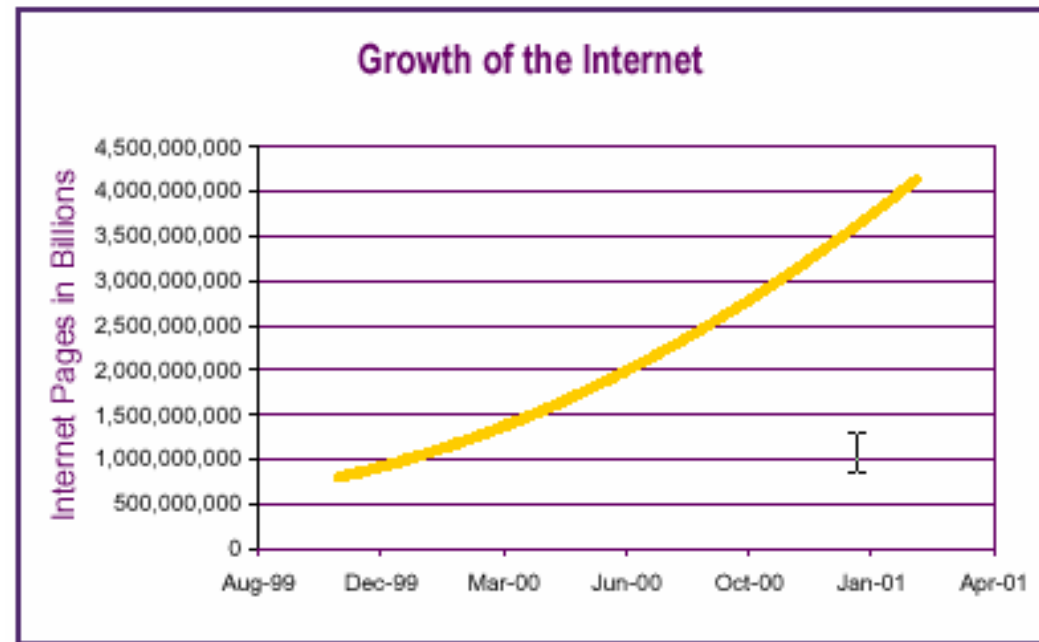
- The US federal government will boost knowledge management spending from \$820 million in 2003 to \$1.3 billion by 2008,
 - largely for homeland security requirements;
 - many European governments are expected to do the same.
- A large chunk of the spending will concern
 - Tools and systems to manage content of unstructured documents

Technologies: Web and KM

- Companies are more and more using the Web for KM
 - The WWW is used as source of information
 - Internal intranet organised as mini Web
 - HTML pages
 - Hyperlinking
 - Search engines used for retrieval
 - Of internal documentation
 - Of external documentation

Web Size Vs Intranet Size

- Web Size: some billion pages (8-???)
- Average Intranet of Large Company:
 - Some dozen million pages
- How long before they reach 1 billion?
 - 2008?
- Web Technologies expected to be key to KM problems
 - What role for the Semantic Web

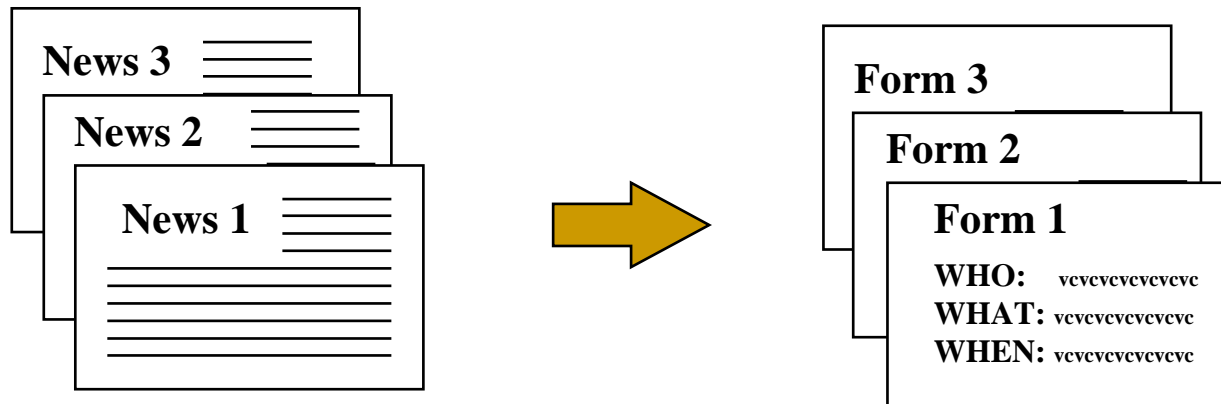


IE from Documents

What Technology

- Information Extraction from Documents
 - Definition
 - Anatomy of a classic IE system (side notes)
- Automated Annotation Using HLT
 - Supporting User Centred Annotation with IE
 - Unsupervised Annotation with IE and Information Integration
 - Adding Knowledge to Documents

Information Extraction



- automatically extracting pre-specified information from natural language texts
 - salient facts about pre-specified types of events, entities or relationships.
- populating a structured information source from a semi-structured, unstructured, or free text, information source.

Standard IE tasks

WASHINGTON, D.C. (October 5, 1999) -
nQuest Inc. today announced that **Paul Jacobs**, for
Vice-President of E-Commerce at **SRA International**
has joined the company's executive management team
as president.

Company: nQuest Inc.

Date: today

InPerson: Paul Jacobs

InRole: president

Company: SRA International

OutPerson: Paul Jacobs

OutRole: Vice-President of E-Commerce,

Named Entities

Event Recognition

Growing complexity

WIT

nlp
sheffield

The generic IE system *[Hobbs 1993]*

- Text Zoner
 - turns a text into a set of text segments (title, body, etc.)
- Preprocessor
 - from a text segment into sequence of sentences
 - morphological analysis
- Filter
 - filters out irrelevant sentences/texts

continue...

Text Zoning

R-15oct93/COMP62

Ident

Title

19:16 Moody's rates Province of Saskatchewan
A3

NEW YORK, Oct 14 (Reuter)

Body

Moody's Investors Service Inc said it assigned an A3 rating to the Province of Saskatchewan's C\$115 million bond offering that was priced today.

The sale is a reopening of the province's 9.6 percent bonds due February 4, 2022.

Proceeds will be used for government purposes, mainly Saskatchewan Power Corp.

Morphological Analysis

19:16 Moody's rates Province of Saskatchewan A3

Say
-Verb
- Past

Moody's Investors Service Inc said it assigned an A3 rating to the Province of Saskatchewan's C\$115 million bond offering that was priced today.

The sale is a reopening of the province's 9.6 percent bonds due February 4, 2022. Proceeds will be used for government purposes, mainly Saskatchewan Power Corp.

Filtering

19:16 Moody's rates Province of Saskatchewan A3

Rating

Local Gvt.

Rating Agency

Moody's Investors Service Inc said it assigned an A3 rating to the Province of Saskatchewan's C\$115 million bond offering that was priced today.

Bond Issue

The sale is a reopening of the province's 9.6 percent bonds due February 4, 2022. Proceeds will be used for government purposes, mainly Saskatchewan Power Corp.

Local_Gvt_Rating

Bond_Issue_Local_Gvt

The generic IE system (contd.)

- **Named Entity Recognizer**
 - identifies small scalable structures (proper names, dates, numbers, currencies, etc.)
- **Parser**
 - produces (possibly complete) parse trees
- **Semantic Interpreter**
 - generates logical forms (LF) for the sentences
- **Lexical Disambiguation**
 - from ambiguous LF to unambiguous LF

continua...

NE Recognition & Coreference

Organisation

19:15 **Moody's** rates Province of Saskatchewan A3

Moody's Investors Service Inc said it assigned an A3 rating to the Province of Saskatchewan's **C\$115 million** bond offering that was priced today.

The sale is a reopening of the province's **9.6 percent** bonds due **February 4, 2022**. Proceeds will be used for **%** government **uses**, mainly Saskatchewan Power Corp.

MNY

Date

The generic IE system (contd.)

- Coreference Resolution
 - identifies different description of the same entity in the text
- Template Generator
 - turns LF into Templates

Template Filling

19:16 Moody's rates Province of Saskatchewan A3

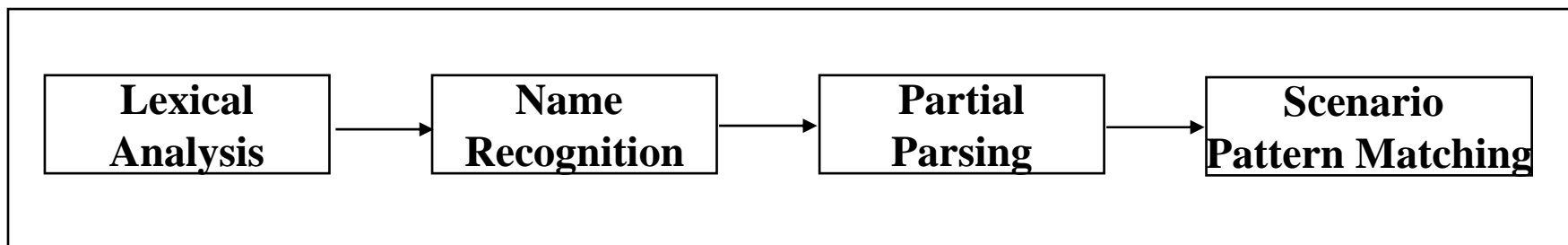
Moody's Investors Service Inc said it assigned an A3 rating to the Province of Saskatchewan's C\$115 million bond offering that was priced today.

The sale is a reopening of the province's 9.6 percent bonds due February 4, 2022. Proceeds will be used for government purposes, mainly Saskatchewan Power Corp.

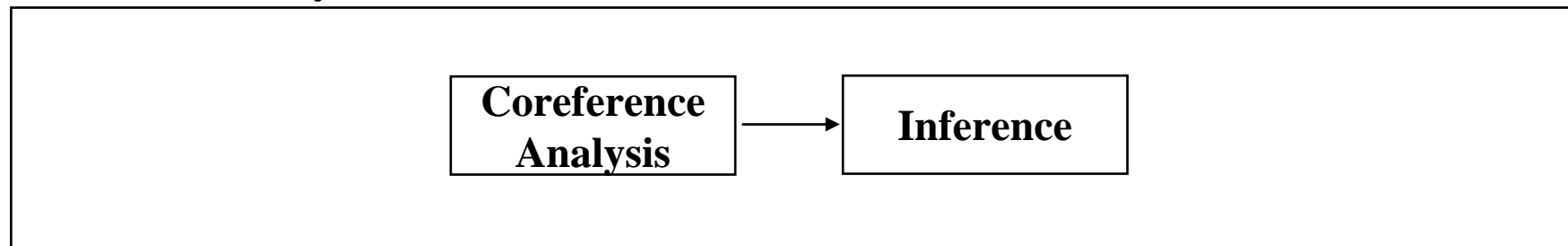
amount	C\$115 million
issuer	Province of Saskatchewan
placement-date	today
maturity	February 4, 2022
rate	9.6 percent

NYU Architecture [Grishman 97]

Local Text Analysis



Discourse Analysis



↓

Template Generation

NYU: Proteus System

Initial Text

Sam Schwarz retired as executive vice president of the famous hot dog manufacturer, Hupplewhite inc. He will be succeeded by Harry Himmelfarb.

EVENT:

Person:

Position:

Company:

NYU: NE Recognition

- Gazetteer lookup

- Patterns:

Person -> FirstName + Word.Capitalised

Person -> Person + Word.Capitalised

Company -> Word.Capitalised+ <company-indicator>

After Name Recognition

[name type: Person Sam Schwarz] retired as
executive vice president of the famous hot
dog manufacturer,

[name type: Company Hupplewhite Inc.] He will be
succeeded by

[name type: Person Harry Himmelfarb].

NYU: Partial Parsing (1)

NP -> (det | indet)? adj* (common | proper)+

VG -> (aux)* verb+

After Partial Parsing(1)

[_{NP} Person e1 Sam Schwarz] [_{VG} retired] as
[_{NP} role e2 executive vice president] of
[_{NP} company e3 the famous hot dog manufacturer],
[_{NP} Company e4 Hupplewhite Inc.] [_{NP} Person e5 He]
[_{VG} will be succeeded] by
[_{NP} Person e6 Harry Himmelfarb].

NYU: Partial Parsing (2)

company -> company-desc comma company-name comma

* action: company(x) ^ hasName(x y) ^ Name(y)

* position -> position of company

action: position(x) ^ position_in(x y) ^ company(y)

[NP Person e1 Sam Schwarz] [VG retired] as
[NP role e2 executive vice president] of
the famous hot dog manufacturer,
Hupplewhite Inc.]
[NP Person e5 He] [VG will be succeeded] by
[NP Person e6 Harry Himmelfarb].

WIT

nlp
sheffield

NYU: Scenario Pattern Matching

clause -> <person> retires as <position>
action: *person(x) ^ leaves_job (x y) ^ job(y)*

clause -> <person> is succeeded by <person>
action: *person(x) ^ succeed (x y) ^ person(y)*

After Scenario Pattern matching

[Clause event e7 Sam Schwarz retired as executive vice president of the famous hot dog manufacturer, Hupplewhite Inc.]

[Clause event e8 He will be succeeded by Harry Himmelfarb]

Coreference

NYU: Final Steps

- Inference

leave-job(x-person, Y-job)
& succeed(Z-person, X-person)
→ start-job(Z-person, Y-job)

- Template Generation

```
EVENT: leave job  
Person: Sam Schwarz  
Position: executive vice president  
Company: Hupplewhite Inc.
```

```
EVENT: start job  
Person: Harry Himmelfarb  
Position: executive vice president  
Company: Hupplewhite Inc.
```

Measures

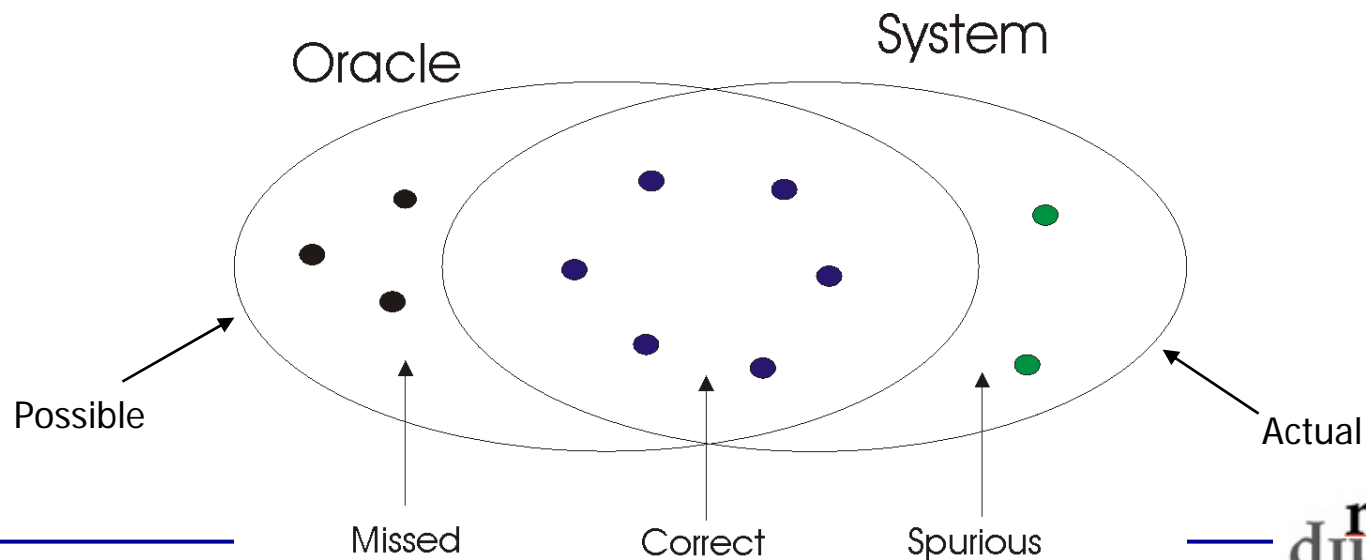
$$\text{Recall} = \frac{\text{CORRECT} + (\text{PARTIAL} * 0.5)}{\text{POSSIBLE}}$$

$$\text{Precision} = \frac{\text{CORRECT} + (\text{PARTIAL} * 0.5)}{\text{ACTUAL}}$$

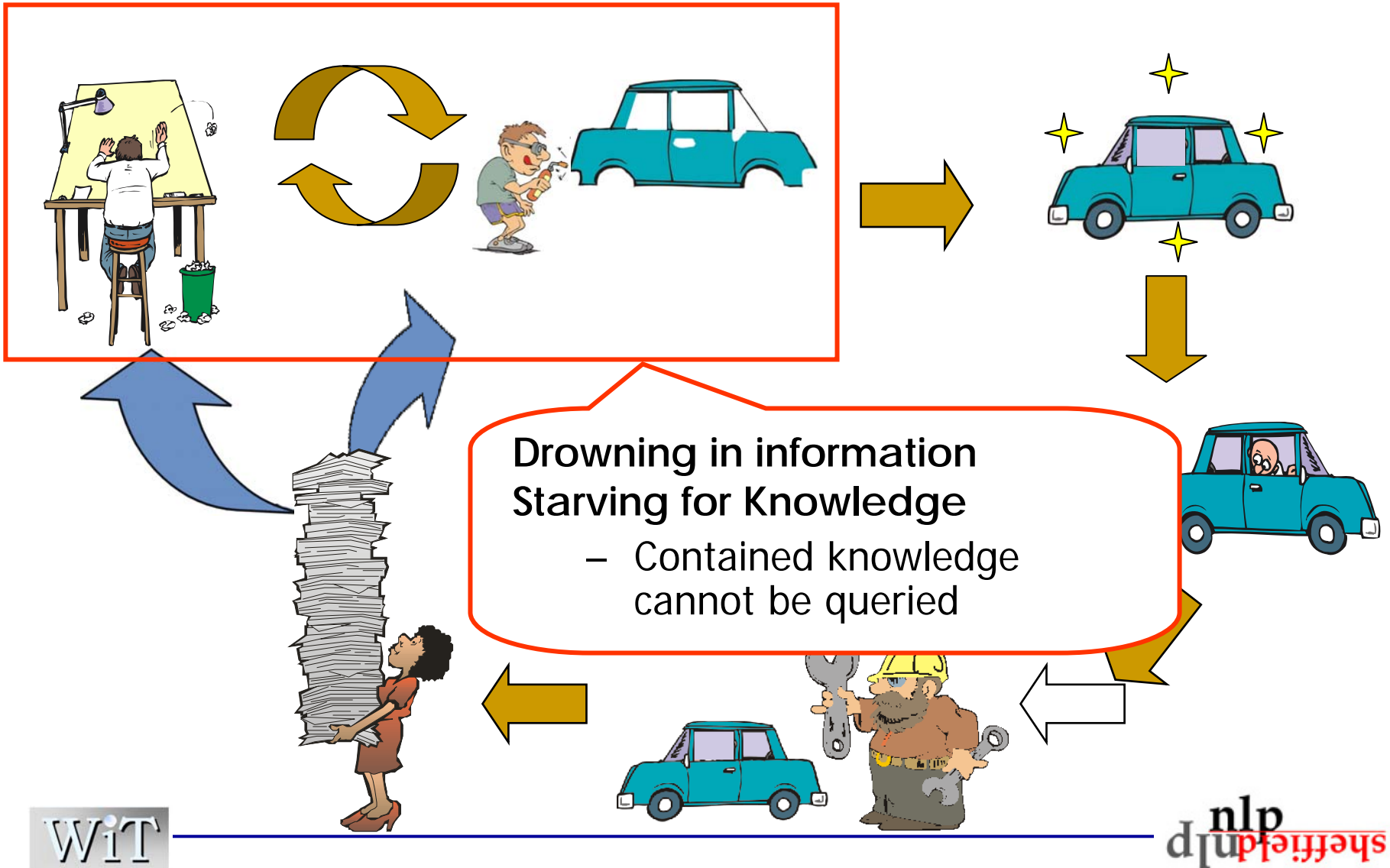
$$F(\beta) = \frac{(\beta^2 + 1) * \text{PREC} * \text{REC}}{\beta^2 * \text{PREC} + \text{REC}}$$

The Rationale Behind

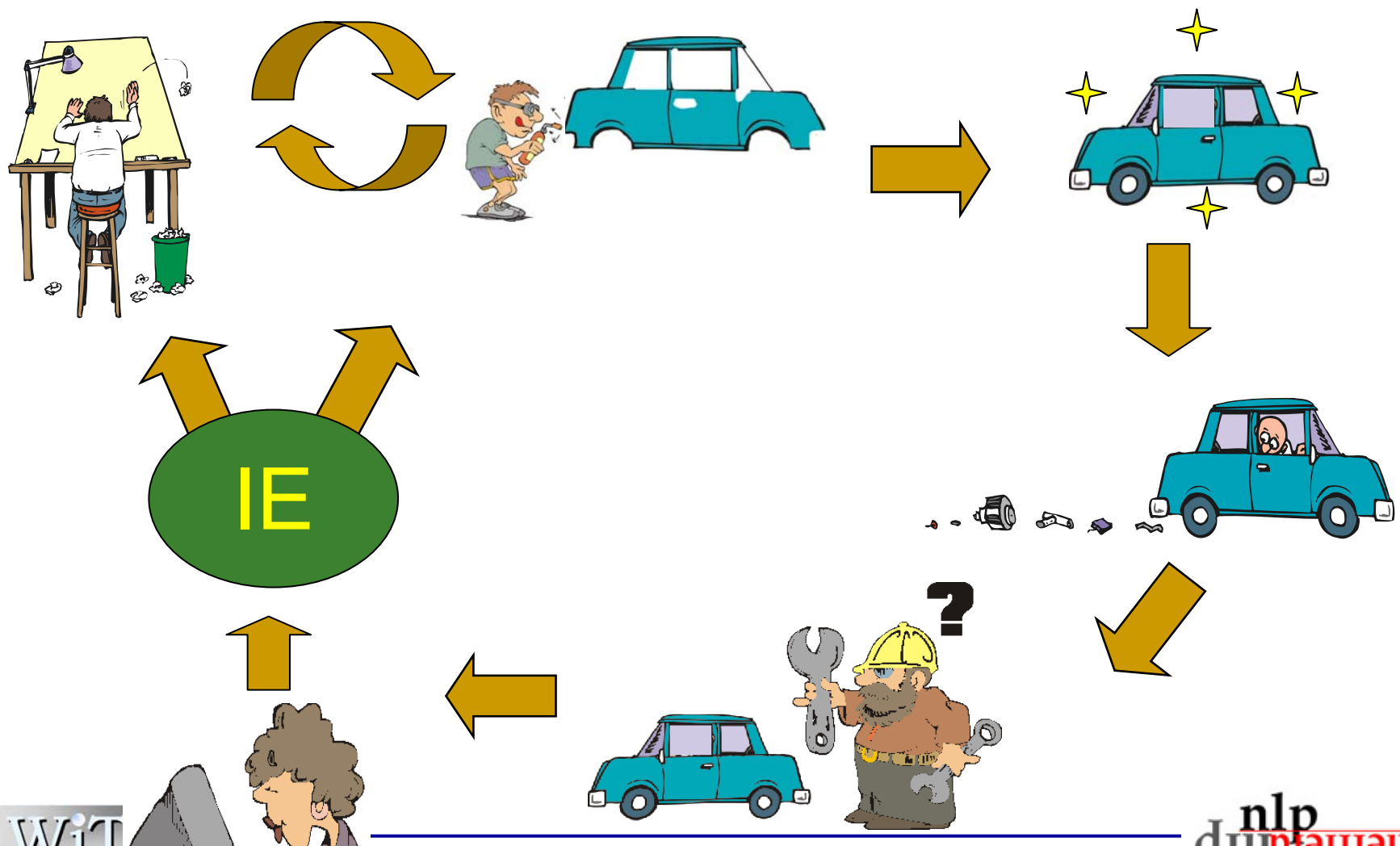
- **Precision:** how correct is the average answer provided by the system
- **Recall:** how many (correct) information are retrieved by the system
- **F-measure:** allows comparative evaluation



Traditional Knowledge Management



Knowledge Management using IE



WIT

Knowledge Management using IE

REF.: 00140/89

STRUCTURED DATA: <licence plate number, model, km,>

TOPIC: Mancato funzionamento motorino avviamento.

TEXT: Sulle auto per presentazione a stampa specializzata si verifica il mancato funzionamento del motorino avviamento durante prova pergola (motorino EY8 0, 8/72).

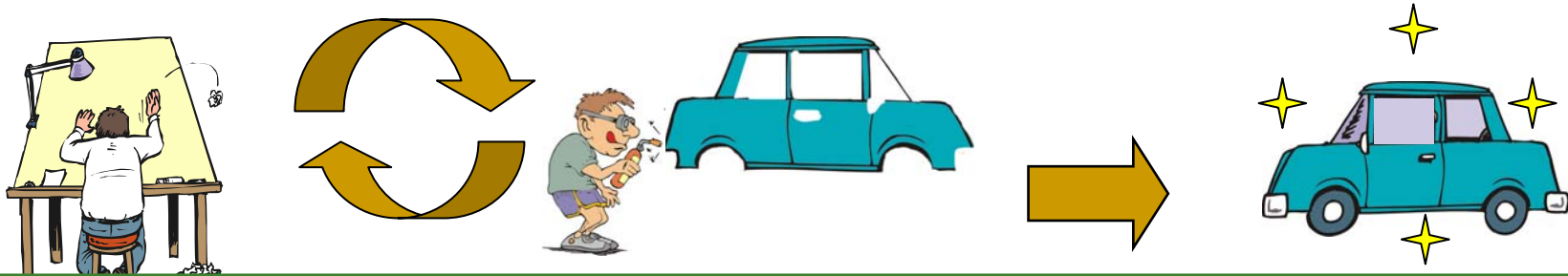
FIRST DIAGNOSIS: Antonioli 24/06/89: vedere scheda 0014/89.

DIAGNOSIS: Bianchi 25/06/89: Anomalia causata da ossidazione con conseguente bloccaggio innesto alberino scorrimento, e mancata chiusura contatti elettromagnete. Il particolare è stato inviato ai laboratori per ulteriori controlli.

Giorgioni 28/06/89 l'ossidazione e' stata causata dall'utilizzo di materiale non idoneo alle prescrizioni.



Knowledge Management using IE



MAIN FAULT: NON-FUNCTIONAL (COD. A124)

Part: starter motor (cod: 0129AIX2)

CAUSED BY: FAILURE TO CLOSE (COD. A156)

Part: electromagnetic contacts starter motor (cod 0129OOT9)

CAUSED BY: BLOCKAGE (COD A345)

Part: starter drive pinion (cod. 0129OOT9)

CAUSED BY: OXIDATION (COD A567)

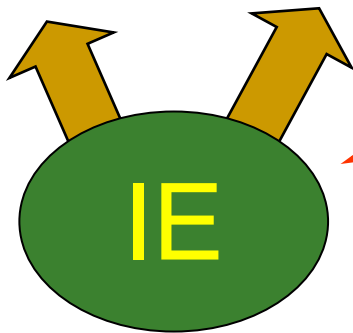
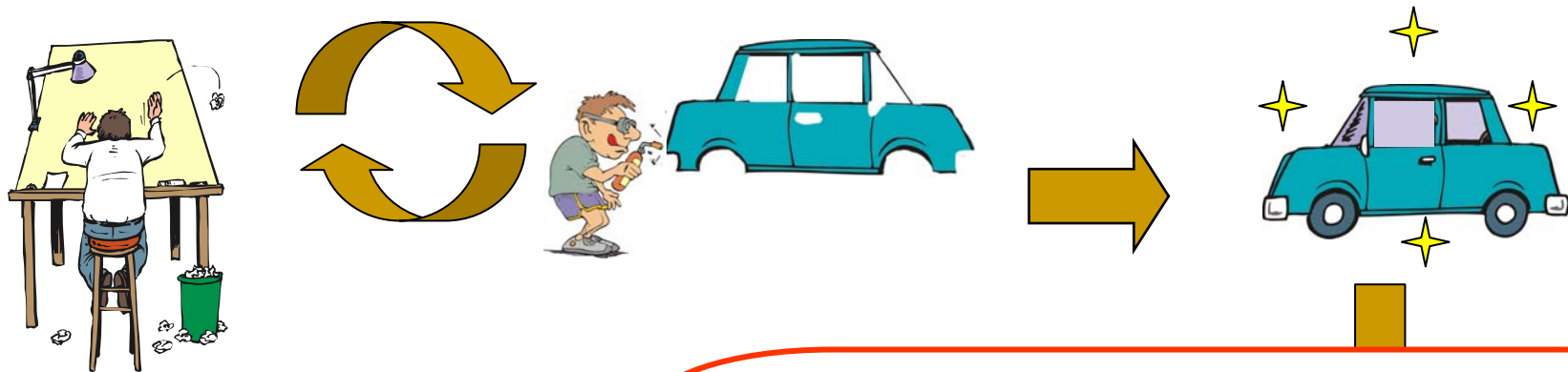
Part: starter drive pinion (cod. 0129OOT9)

CAUSED BY: UNSUITABLE MATERIAL (COD A569)

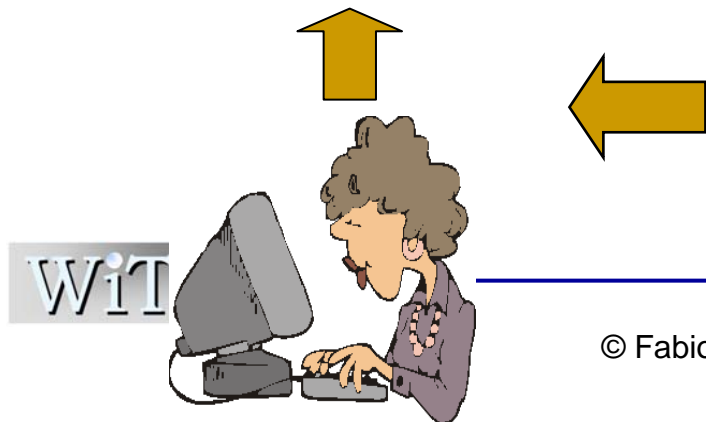
Part: starter drive pinion (cod. 0129OOT9)



Knowledge Management using IE



- Direct access to knowledge
- Speed: Prompt Identification of critical factors
- Quality: only relevant information retrieved
- Quantity: more information can be accessed
- Knowledge Sharing



WIT

IE Tools: a very partial list

- Requiring manual development

- Fastus (SRI)
- Lasie (Ushf)
- Proteus (NYU)
- Annie (Ushf, www.gate.ac.uk)
- ...

- Based on Machine Learning

- Alembic (Mitre, www.mitre.org/tech/alembic-workbench/)
- SIFT (BBN)
- Amilcare (Ushf, nlp.shef.ac.uk/amilcare/)
- ...

- A General Architecture for Text Engineering:
architecture, framework

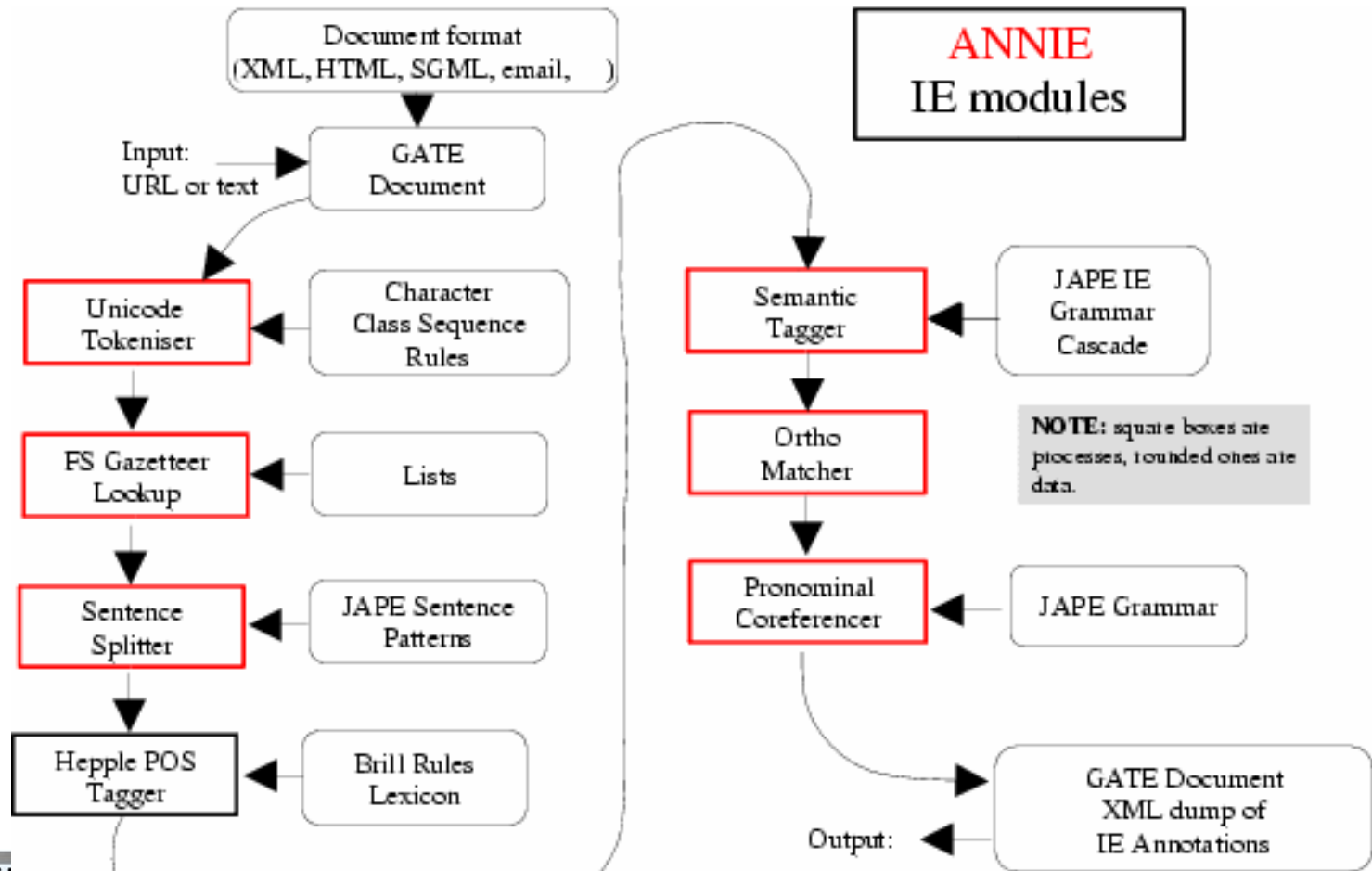
Why?

- Free software, relatively comprehensive, widely used,
- It means we can ignore the infrastructural issues
- Not a claim that it is the best or only in all cases!

Gate- Annie

- ANNIE – A Nearly-New IE system
- A version distributed as part of GATE
- GATE automatically deals with document formats, saving of results, evaluation, and visualisation of results for debugging
- GATE has a finite-state pattern-action rule language, used by ANNIE
- A reusable and easily extendable set of components

NE Components



On Image Annotation

Annotating Images

- Images do not have content like text
- Can be annotated by
 - Selecting regions
 - Associating annotations to regions
 - In a way similar to CREAM



Motor Cycle Cop

photo by [ph](#) [PoliceMan.rdf](#) [Full size](#)

Add a comment, cirave:



add comment



WWW Conference Photos

annotate photo

PoliceMan.jpg



Motor Cycle Cop

photo by [ph](#) [PoliceMan.rdf](#) [Full size](#)

Logged in as cirave

- [My account settings](#)
- [Customize person list](#)
- [Show my photos](#)
- [Logout](#)

You are in Gallery "www2004"

- [Show published photos](#)
- [Vote - great shot or not?](#)
- [Add photos](#)

Select...

Add annotation



If this is a person,

Select name [Customize person list](#)

Person's email (A quick hack for Dev Day Demo - under development)

Annotation headline

Additional comment or description



Annotated by: cirave

Save

Cancel

- This is a very simple example of how to annotate photos
 - Limitations
 - Ontology is very limited (just one concept)
 - Interesting issues
 - A community building the SW
 - Sharing of knowledge

MIAKT (2002-2004)

- Support to the Multi-Disciplinary Meetings (MDMs) that take place between various medical practitioners of different expertise, in coming to a collaborative diagnosis and plan of action in symptomatic focal breast disease.

MIAKT: Multi-disciplinary Assessment

- Multiple stakeholders
- Multiple viewpoints and vocabularies
 - Breast imaging – X-ray, ultrasound, MRI
 - Clinical examination
 - Microscopy – cells and tissues (also, hormone receptors)
- Local dialects in use
- Variation between countries due to factors such as insurance claims!

