

An Information Extraction Perspective on Text Mining: Tasks, Technologies and Prototype Applications

Robert Gaizauskas

*Natural Language Processing Group
Department of Computer Science,*



University of Sheffield

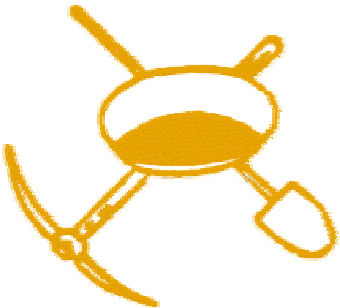
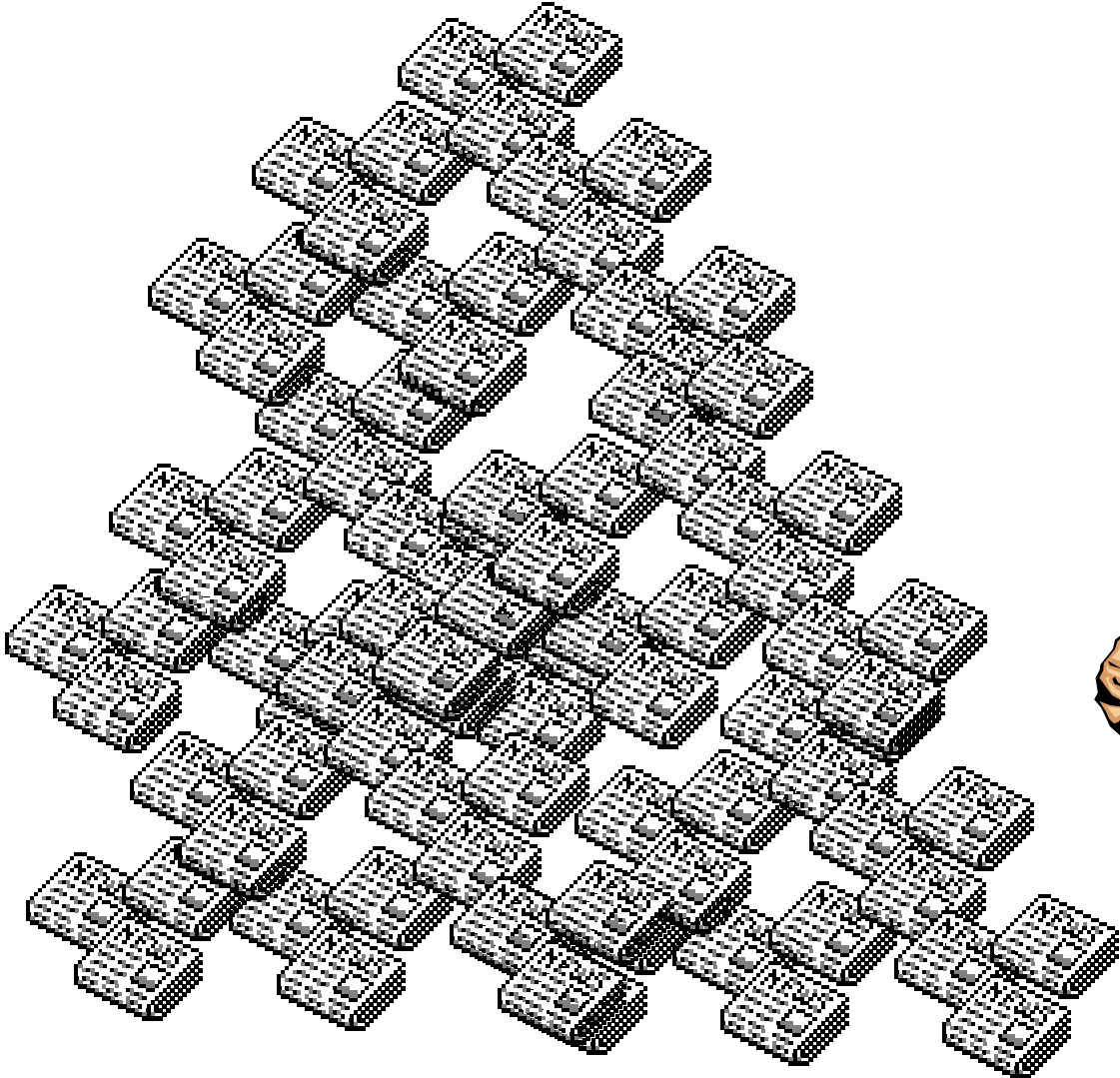
Outline of Talk

- The Text Mining Scenario
- Information Extraction: Definition and Scope
- Information Extraction: Component Tasks
- Information Extraction: Technologies
- Information Extraction: Prototype Application
- Conclusions and Future Directions/Challenges

Outline of Talk

- The Text Mining Scenario
- Information Extraction: Definition and Scope
- Information Extraction: Component Tasks
- Information Extraction: Technologies
- Information Extraction: Prototype Application
- Conclusions and Future Directions/Challenges

Text Mining: Scenario



Text Mining Scenario Components: Texts

■ Genres

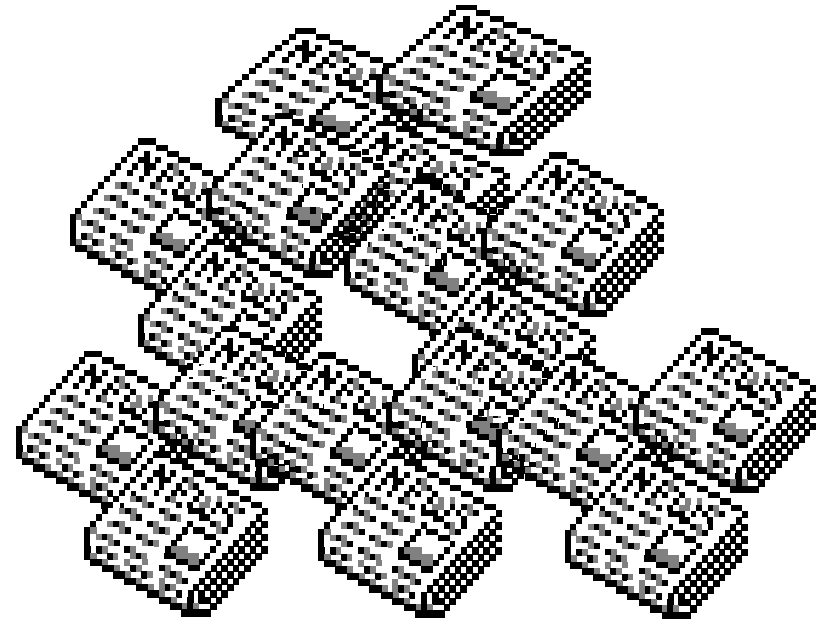
- Newspapers
- Company reports
- Web pages
- Scientific papers
- Legal documents

■ E-Formats

- Word Documents
- PDF/Postscript
- HTML/SGML/XML

■ Languages

- English ... French ... Greek ... Russian ... Chinese ... Hindi ... Sanskrit
... Linear B
- Character encodings: ASCII, ISO 8859, Unicode, ...



Text Mining Scenario Components: Users

- User domain of interest
 - Business – competitor intelligence, corporate intranet/memory
 - Scientists – access to literature
 - Military/police intelligence – open source intelligence, intranet
 - Journalists – news archives
- User level of domain expertise
 - Novice/expert
- User linguistic competence
 - Adult/child
 - Native/non-native language speaker
 - Uni/multi-lingual



Text Mining Scenario Components: Information Access Needs

■ Ad hoc searching

- Specific questions: *What year did the Berlin Wall come down?*
- General background/context: *Tell me about Zakopane*



■ Stable intelligence gathering

- Scenario-related: *Build a database recording new projects in the energy sector: the players, location, energy type, start date, capitalisation*
- Entity-related: *Build a database of key scientists in the biotech industry: name, employer, position, start and end dates*



■ Current awareness

- Alerting: *Let me know when any papers are published on the crystallographic structure of any lipase*
- Document selection: *Assemble articles on drug approvals*



Text Mining Scenario Components: Information Access Needs

■ Summarisation

- Single/multi-document: *Summarise the Bulger trial*



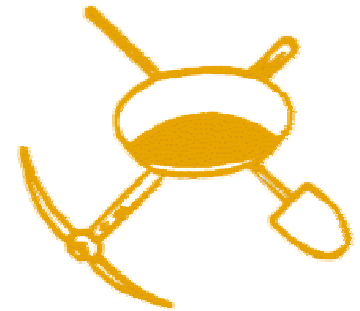
■ Knowledge discovery

- Trends/correlations in time series data, e.g. commodity price changes
- Transitive linkages, e.g between businesses, people, enzymes



Text Mining Scenario Components: Tools

- Information retrieval ←
- Document clustering/routing/categorisation
- Information extraction ←
- Summarisation
- Agents
- Web crawlers
- KDD/data mining



Information Extraction is a necessary, though not sufficient component of any text **content** mining scenario

Outline of Talk

- The Text Mining Scenario
- Information Extraction – Definition and Scope
 - Definition + Example
 - Contrast with Information Retrieval
 - Challenges, Strengths, Weaknesses, Appropriate Use
 - Brief History
 - Evaluation Methodology
- Information Extraction Component Tasks
- Information Extraction Technologies
- Information Extraction Prototype Applications
- Conclusions and Future Directions/Challenges

What is Information Extraction?

- The Information Extraction (IE) task: from each text in a set of natural language texts extract information about predefined classes of entities and relationships and place this information into a **template** or database record

E.g. from financial newswire stories identify those dealing with management succession events and extract from them details of organisations and persons, the post being assumed or vacated, the reason for vacancy, etc.

- IE may also be described as the activity of populating a structured information repository (database) from an unstructured, or free text, information source

What is Information Extraction? (cont)

The resulting structured database is then used for some other purpose:

- searching or analysis using conventional database queries;
- data-mining;
- generating a summary (perhaps in another language);
- constructing indices into/within/between the source texts.

Example: A *Wall Street Journal* Article

<DOC>

<DOCID> wsj94_008.0212 </DOCID>

<DOCNO> 940413-0062. </DOCNO>

<HL> Who's News:

@ Burns Fry Ltd. </HL>

<DD> 04/13/94 </DD>

<SO> WALL STREET JOURNAL (J), PAGE B10 </SO>

<CO> MER </CO>

<IN> SECURITIES (SCR) </IN>

<TXT>

<p>

BURNS FRY Ltd. (Toronto) -- Donald Wright, 46 years old, was named executive vice president and director of fixed income at this brokerage firm. Mr. Wright resigned as president of Merrill Lynch Canada Inc., a unit of Merrill Lynch & Co., to succeed Mark Kassirer, 48, who left Burns Fry last month. A Merrill Lynch spokeswoman said it hasn't named a successor to Mr. Wright, who is expected to begin his new position by the end of the month.

</p>

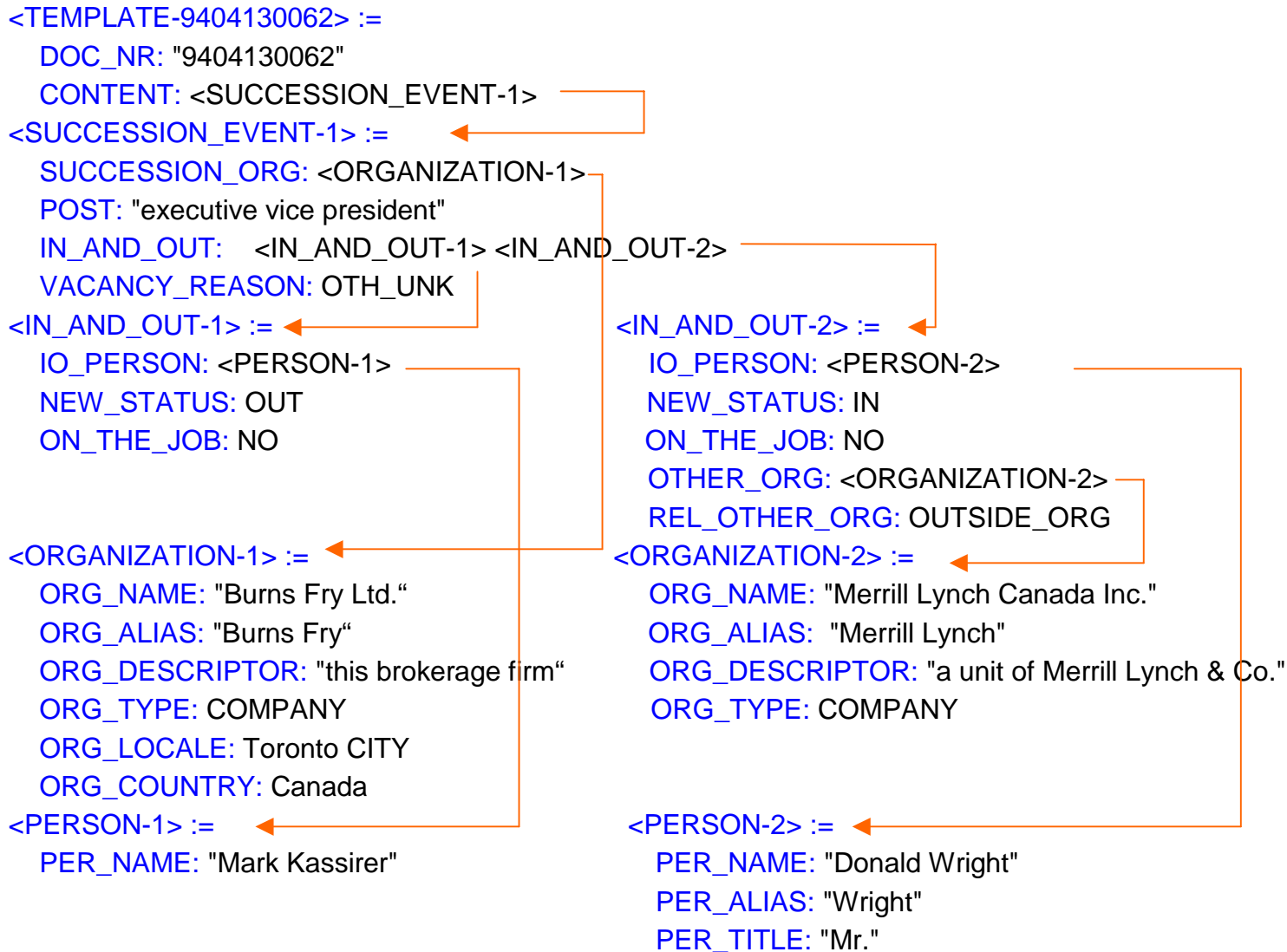
</TXT>

</DOC>

Example: Management Succession Event Template Definition

```
<TEMPLATE> :=
  DOC_NR:          "NUMBER" ^
  CONTENT:         <SUCCESSION_EVENT> *
<SUCCESSION_EVENT> :=
  ORGANIZATION:   <ORGANIZATION> ^
  POST:           "POSITION TITLE" | "no title" ^
  IN_AND_OUT:     <IN_AND_OUT> +
  VACANCY_REASON: {DEPART_WORKFORCE, REASSIGNMENT, NEW_POST_CREATED, OTH_UNK} ^
<IN_AND_OUT> :=
  PERSON:         <PERSON> ^
  NEW_STATUS:     {IN, IN_ACTING, OUT, OUT_ACTING} ^
  ON_THE_JOB:     {YES, NO, UNCLEAR}
  OTHER_ORG:      <ORGANIZATION> -
  REL_OTHER_ORG:  {SAME_ORG, RELATED_ORG, OUTSIDE_ORG} -
<ORGANIZATION> :=
  ORG_NAME:       "NAME" -
  ORG_ALIAS:      "ALIAS" *
  ORG_DESCRIPTOR: "DESCRIPTOR" -
  ORG_TYPE:       {GOVERNMENT, COMPANY, OTHER} ^
  ORG_LOCALE:     LOCALE_STRING {{CITY, PROVINCE, COUNTRY, REGION, UNK} *
  ORG_COUNTRY:    NORMALIZED-COUNTRY-or-REGION | COUNTRY-or-REGION-STRING *
<PERSON> :=
  PER_NAME:       "NAME" -
  PER_ALIAS:      "ALIAS" *
  PER_TITLE:      "TITLE" *
```

Example: Filled Management Succession Event Template



Example: Uses for Templates

- From the completely filled version of the preceding template a natural language summary can be generated:

BURNS FRY Ltd. named Donald Wright as executive vice president.
Donald Wright resigned as president of Merrill Lynch Canada Inc..
Mark Kassirer left as president of BURNS FRY Ltd.

- Or, a table can be constructed:.

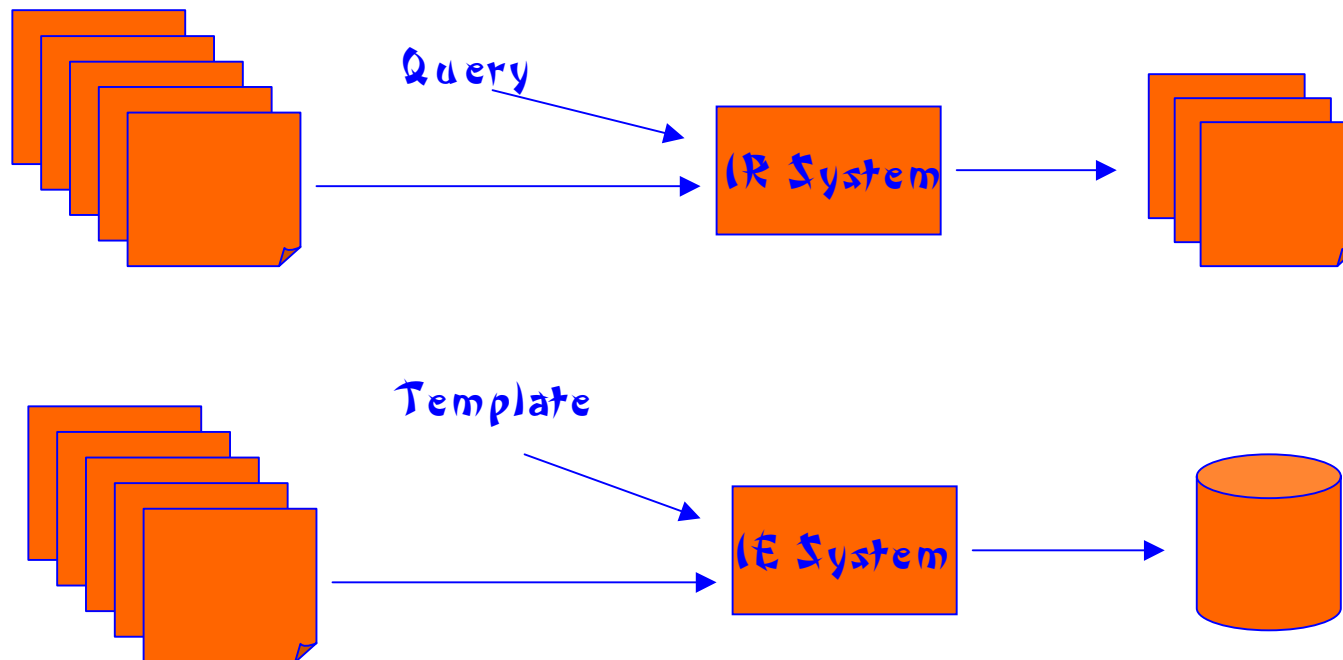
Company	Post	Person	Direction
Burns Fry	Executive VP	Donald Wright	In
Burns Fry	President	Mark Kassirer	Out
Merrill Lynch Canada	President	Donald Wright	Out

What IE is **NOT**: Information Retrieval

- The Information Retrieval (IR) task: given a user query and a document collection retrieve that subset of documents from the collection which are relevant to the user's query.
- E.g. given the query
chief executive officer head president chairman post succeed name
return those documents in last year's *Wall Street Journal* pertaining to management succession events.
- Once the IR system returns the documents, the user browses the selected documents in order to fulfil his or her information need.
- Depending on the IR system, the user may be further assisted by
 - relevance ranking of retrieved documents
 - highlighting of search terms in the text to facilitate identifying passages of particular interest

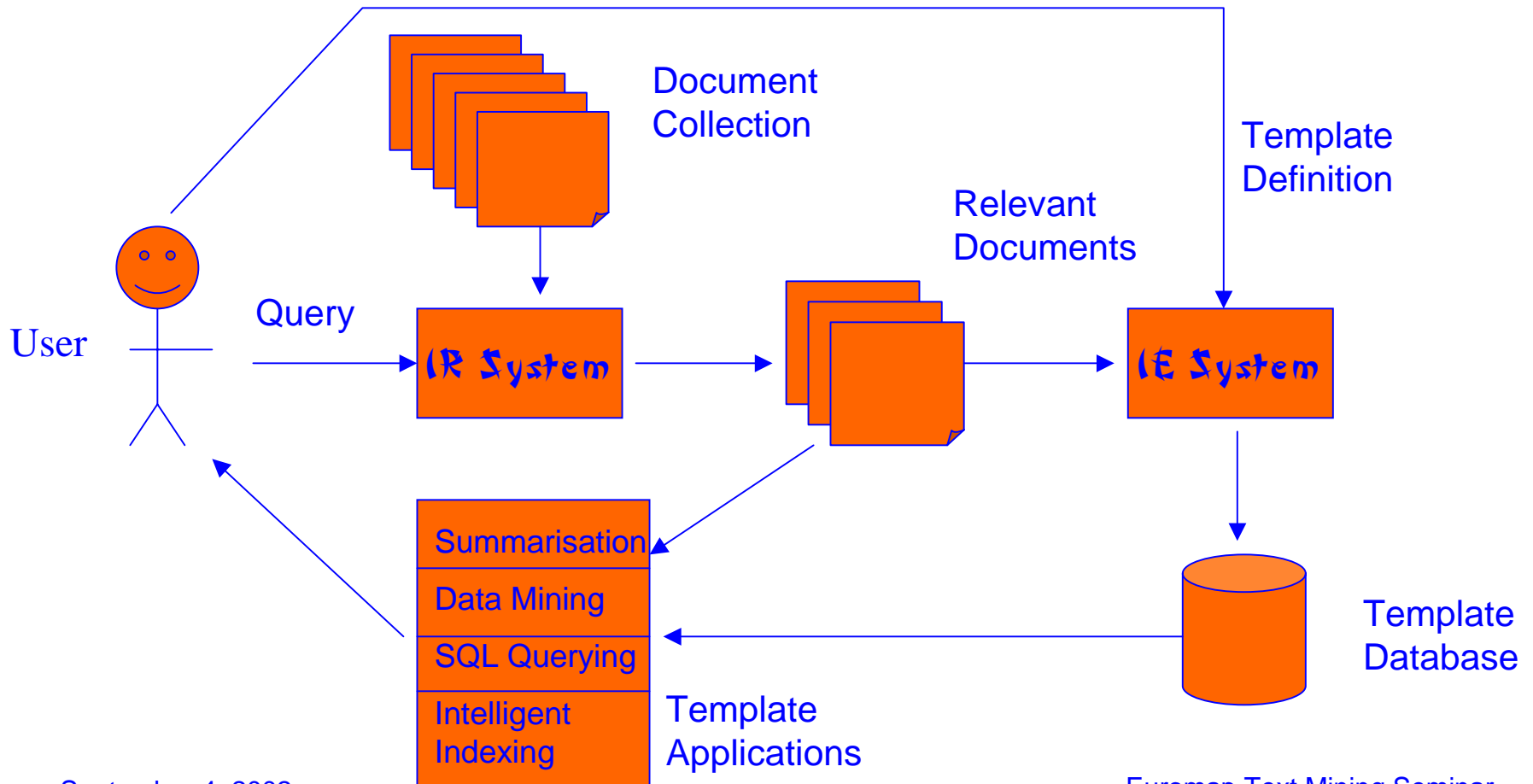
The Contrast Between IR and IE: Goals

- IR retrieves relevant documents from collections
- IE extracts relevant information from documents
- IR delivers documents to the user
- IE delivers facts to the user/other applications



Integrating IR and IE

- IR and IE are not in competition -- they are complementary and their use in combination has the potential to create powerful new tools in text processing.



Why IE is Hard

- Because natural language is hard ...
- Language is **flexible** – there are many ways of saying the same thing
 1. Gina Torretta succeeds Nicholas Andrews as chairperson of BNC Holdings Inc.
 2. BNC Holdings Inc. named Ms G. Torretta as its new chair-person after Nick Andrews' departure
 3. Nicholas Andrews was succeeded by Gina Torretta as chair-person of BNC Holdings
- Language is **ambiguous** – the same way of saying something can mean different things
 1. They bought the company with last year's profits
 2. They bought the company with 1000 employees
- Language is **dynamic**
 - New words are constantly appearing: ecotourist, ramraider, dis, prozac
 - Old words gain new senses: to text

Strengths and Weaknesses of IE

Strengths:

- Extracts facts from texts, not just texts from text collections
- Can feed other powerful applications (databases, indexing engines)

Weaknesses:

- Porting to new genres and domains is time-consuming and requires expertise
- Limited accuracy
- Not fast enough to run over large text collections while user waits

Strengths and Weaknesses of IR

Strengths:

- Can search huge document collections very rapidly
- Insensitive to genre and domain of the texts
- Can rank documents with respect to likely relevance
- Searches can be iteratively refined

Weaknesses:

- Documents are returned not information/answers, so
 - user must further read texts to extract information
 - output cannot be directly data mined
- Frequently not discriminating enough (“1563 documents match your request”)

A Brief History of IE

- The first published work on information extraction (though it was not called this at the time) was in late 1960s
 - A significant precursor was the psychologist Roger Schank's work on scripts and story understanding in the 1970's
 - The 1980's saw the emergence of some commercial systems targetted at financial transactions and newswires
 - The big impetus to current research started in the late 1980's when DARPA initiated a series of competitive evaluations of "Message Understanding" systems (Message Understanding Conferences – MUC)
- MUC ran for 10 years (1987-98) and significantly advanced the field
- Currently there are a number of IE systems on the market and a large and on-going research effort in the field

Outline of Talk

- The Text Mining Scenario
- Information Extraction: Definition and Scope
- Information Extraction Component Tasks
 - Entity Extraction
 - Attribute Extraction
 - Relation Extraction
 - Event Extraction
- Information Extraction: Technologies
- Information Extraction: Prototype Applications
- Conclusions and Future Directions/Challenges

IE Component Tasks

- To fill complex templates IE researchers have discovered that systems must be able to perform a variety of simpler tasks
- Studying and evaluating these component tasks in isolation has proved a useful way forward for IE research
- Simpler tasks may also prove sufficient/useful for applications in their own right

Entity Extraction

- Types of entities which have been addressed by IE systems:
 - Named individuals
 - Organisations, persons, locations, books, films, ships, restaurants, hotels, ...
 - Named kinds
 - Proteins, chemical compounds, drugs, diseases, aircraft components, ...
 - Times
 - temporal expressions – dates, times of day
 - Measures
 - monetary expressions, distances/sizes, weights ...

- For each textual reference must identify its extent and its type

Cable and Wireless today announced

IBM and Microsoft today announced

= company

John Lewis ...

= company or person?

Entity Extraction: Coreference (1)

- Multiple references to the same entity in a text are rarely made using the same string:
 - Pronouns – Tony Blair ... he
 - Names/definite descriptions – Tony Blair ... the Prime Minister
 - Abbreviations/acronyms – Tony Blair ... Blair; United Nations ... UN
 - Orthographic variants – alpha helix ... alpha-helix ... a-helix ... A-helix
- IE systems more useful if they can link all references to the same entity

Attribute Extraction

- Entities frequently have associated **attributes** of interest

E.g.

<ORGANIZATION-1> :=

ORG_NAME: "Burns Fry Ltd."

ORG_ALIAS: "Burns Fry"

ORG_DESCRIPTOR: "this brokerage firm"

ORG_TYPE: COMPANY

ORG_LOCALE: Toronto CITY

ORG_COUNTRY: Canada

BURNS FRY Ltd. (Toronto) -- Donald Wright, 46 years old, was named executive vice president and director of fixed income at this brokerage firm. Mr. Wright resigned as president of Merrill Lynch Canada Inc., a unit of Merrill Lynch & Co., to succeed Mark Kassirer, 48, who left Burns Fry last month. A Merrill Lynch spokeswoman said it hasn't named a successor to Mr. Wright, who is expected to begin his new position by the end of the month.

Attribute Extraction: Coreference (2)

- Discovering attribute values frequently depends upon being able to follow coreference links.

Dirk Ruthless of MegaCorp made a stunning announcement today. In September he will be stepping down as Chief Executive Officer to spend more time with his pet piranhas

To determine the corporate position of Dirk Ruthless from we must correctly resolve the pronominal anaphor “he” in the second sentence with “Dirk Ruthless” in the first.

Relation Extraction

- Given the ability to extract entities, and associated attributes, the next step is to extract **relations** holding between these entities.

E.g.

- EMPLOYEE_OF holds between PERSON and ORGANISATION
- PRODUCT_OF holds between ARTIFACT and ORGANISATION
- IN_PROTEIN holds between RESIDUES and PROTEINS

Event Extraction

- Many domains are characterised by key events or scenarios
- Example scenarios:
 - Personnel movement events within organisations
 - Joint venture announcements
 - Product announcements
 - Transportation disasters
 - Terrorist attacks
 - Enzyme interactions
- Events can be viewed as complex relations, whose position in time is typically of key importance

IE Evaluation Methodology

- Correct answers, called **keys**, are produced manually for each extraction task (filled templates or SGML annotated texts)
- Scoring of system results, called **responses**, against keys is done automatically.
- Some portion of the answer keys are multiply produced by different humans so that **interannotator agreement** figures can be computed.
- Principal metrics are:
 - **Precision** (how much of what system returns is correct)
 - **Recall** (how much of what is correct system returns)
 - **F-measure** (a weighted combination of precision and recall)
- Objective evaluation metrics are essential for
 - Inter-system comparison
 - Regression testing
 - Machine learning

State-of-the-art Evaluation Results in Open International Competition (MUC-7, 1998)

Task	Recall	Precision	P & R
Named Entity	92	95	93.39
Coreference	56.1	68.8	61.8
Template Element (entity + attributes)	86	87	86.76
Template Relation	67	86	75.63
Scenario Template	42	65	50.79

Human performance ranges from ~95% down to ~80%, depending on task

Outline of Talk

- The Text Mining Scenario
- Information Extraction Tasks + Methodologies
- Information Extraction Component Tasks
- Information Extraction Technologies
 - Linguistically-motivated IE (aka “deep” IE)
 - Cascaded finite state transducers (aka “shallow” IE)
 - Adaptivity via Machine learning
- information Extraction Prototype Applications
- Conclusions and Future Directions/Challenges

Linguistically Motivated IE

- During the 1980's and early 90's, NLP dominated by
 - Descriptive linguistics
 - Symbolic AI techniques
- Outcome:
 - Rich computational models of grammar (GPSG, LFG, HPSG, etc.)
 - Logic-based approaches to semantics (Montague)
 - Logic-based approaches to discourse (DRT, abduction)
- Presumption: IE requires, or at least will benefit from, the best computational models of human language processing
- One system which is in this tradition is the **LaSIE** (LArge Scale Information Extraction) system developed at Sheffield ...



The LaSIE System

LaSIE processes texts in four principal stages:

- text preprocessing
- lexical and terminological processing
- syntactic analysis and semantic interpretation
- discourse interpretation

A final **template writing** stage generates template output as required

Text Preprocessing

■ Text structure analysis

- Many text types have regular structure – due to genre and/or publishing conventions
 - Structure may be explicit via markup or formatting
 - E.g. scientific abstracts/papers, newswires, botanical flora, ...
- For particular applications certain sections can be targeted for detailed analysis while others can be skipped completely
- Where texts are marked up in SGML with a DTD, an initial module parses the markup
- Where articles are in plain text, an initial 'sectioniser' module is used to identify and classify significant sections using sets of regular expressions.

■ Tokenisation

- separates input text into tokens – whitespace, symbols, SGML tags, numbers, words (various case distinctions)
- in addition to the normal white-space/punctuation delimited tokenisation required for newswires, scientific papers require further sophistication: NaCl ,Tyr152

Lexical and Terminological Processing

■ Morphological processing

- Standard English inflectional forms are recognised
- Domain-specific morphological rules are used to help recognise terminology (e.g. dehydrogenase)

■ Terminological processing

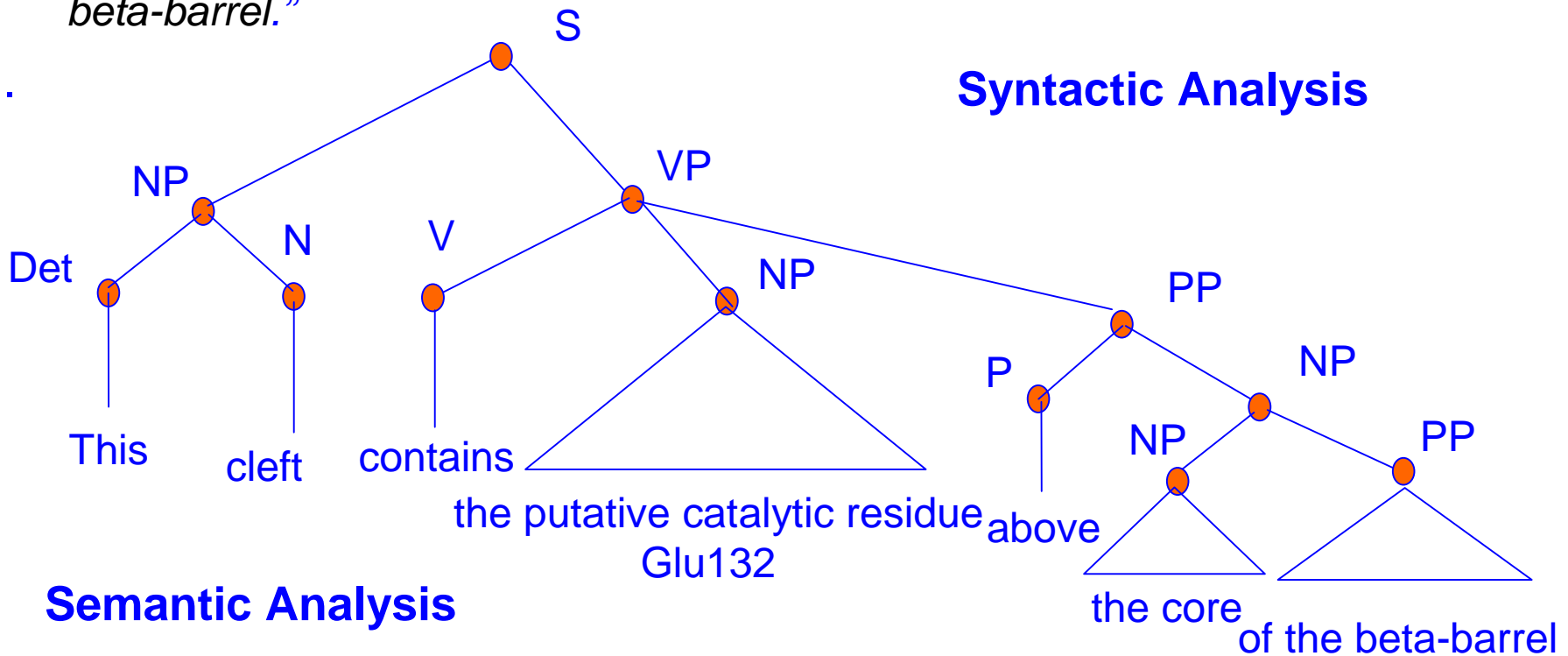
- Multi-token sequences matched against specialised lexicons
 - In business domains: company names, person names, locations ...
 - In biological scientific domains: proteins, residues, enzymes, elements, measures (> 25,000 terms in 52 lexicons)
- Multi-token sequences parsed using specialised grammars
 - In business domains: ORG -> LOC ORG_KEY+ COMP_DES
Norwich Investment Bank plc.
 - In biological scientific domains: mannitol-1-phosphate5-dehydrogenase
 - Grammars give generative capacity (i.e. can recognise new terms)

Syntactic Analysis and Semantic Interpretation

- **Sentence splitting**
 - Break text into sentences (identify sentence-final full stops)
- **Part-of-speech tagging**
 - Assign word class to each word/term (e.g. NN,NNS,VBG)
 - Terms recognised in terminological processing treated as non-decomposable units, with syntactic role of proper noun
- **Syntactic Analysis**
 - Assign (partial) phrase structure using general phrasal grammar of English
 - Grammar represented in feature-based unification formalism
 - Parsing carried out by bottom-up chart parser
- **Semantic Interpretation**
 - Grammar includes compositional semantic rules, which are used to construct a “meaning” representation of the “best” parse of each sentence.
 - This predicate logic-like representation is passed on as input to the discourse interpretation stage.

Syntactic Analysis and Semantic Interpretation: Example

“This cleft contains the putative catalytic residue Glu132 above the core of the beta-barrel.”



contain(e1),
cleft(e2), lsubj(e1,2), det(e2,this),
residue(e3), lobj(e1,e3), name(e3,"Glu132"), adj(e3,putative), adj(e3,catalytic)
core(e4), above(e1,e4)
secondary_structure(e5), name(e5,"beta-barrel"), of(e4,e5)

Discourse Interpretation

- The semantic representation of each sentence is added to a predefined **domain model** made up of
 - an **ontology**, or concept hierarchy, and
 - inheritable **attributes** and **inference rules** associated with concept nodes in the hierarchy
- The domain model is gradually populated with instances of concepts from the text to become a **discourse model**
- A **coreference mechanism** attempts to merge each newly introduced instance with an existing one, subject to various syntactic and semantic constraints
- **Inference rules** associated with particular instance types may hypothesise the existence of further presupposed instances (e.g. a resignation event presupposes an organisation)
- The coreference mechanism then attempts to resolve the hypothesised instances with actual instances from the text
 - models implicit content of text
 - redresses deficiencies in parsing

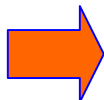
Discourse Interpretation: Example

1. *The three-dimensional structure of Endo H has been determined ...*
2. *A shallow curved cleft runs across the surface of the molecule from ...*
3. *This cleft contains the putative catalytic residues Asp130 and Glu132 ...*

- **In which protein is Glu132?**
- From 1, Endo H is identified as a protein – `protein(e1),name(e1,"Endo H")` – and added to the discourse model
- From 2, the cleft is identified – `cleft(e23)` – and the molecule – `molecule(e25)`
 - Ontology records that proteins are molecules and coreference resolves `e25` and `e1`
 - Domain model/ontology records that clefts are regions and that regions are located in proteins – a protein, say `e42`, is hypothesized and the relation `located_in(e23,e42)`
 - In the absence of full semantic analysis of “runs across the surface of”, coreference picks the closest protein and resolve `e42` with `e1/e25` – i.e. the cleft is assumed to be in EndoH
- From 3, the analysis is as before – the cleft is identified as, say `e52`, and the residue, `e61`
 - coreference resolves the cleft `e52` with the preceding `e23`
 - The domain model allows reasoning from “contains” to establish the relation `located_in(e61,e23)` – the residue is located in the cleft
 - Transitivity of `located_in` permits the conclusion: `located_in(e61,e1)` – **Glu132 is in EndoH**

Cascaded Finite State Transducers

- Linguistically motivated approaches to IE have proved reasonably successful, but have a number of disadvantages
 - Computationally expensive (= slow)
 - Complex (= expensive)
 - Slow to build
 - Require narrow expertise
 - Domain-specific – difficult to move to new domains
- Does IE require rich models of human language processing?
- Consider simpler approaches that
 - are fast
 - only aim at template filling, and not at comprehensive, accurate language processing
 - can be trained on annotated/unannotated text collections (or *corpora*)



cascaded finite state transducers

Cascaded Finite State Transducers (cont)

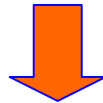
- Finite state transducers process a sequential input and, if it meets certain conditions, emit output dependent only on the current input and state of the device
- FSTs are simple, highly efficient computational devices that can be used to recognise patterns in text and “transduce”, i.e. convert, the input stream to a modified output stream
- Multiple FSTs can be run in sequence (“cascaded”) each of which performs a simple language processing task
- Result may be an adequate extraction system, though it makes no attempt to model a general linguistic processing

Cascaded Finite State Transducers (cont)

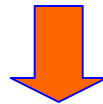
Gina Torretta succeeds Nicholas Andrews as
chairperson of BNC Holdings Inc



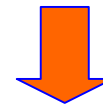
<PERSON> succeeds <PERSON> as
chairperson of BNC Holdings Inc



<PERSON> succeeds <PERSON> as
chairperson of <COMPANY>



<PERSON> succeeds <PERSON> as
<POST> of <COMPANY>



Template

Person Name
FST

Company Name
FST

Post Name
FST

Scenario
FST

Cascaded Finite State Transducers (cont)

- IE systems based on cascaded FSTs have become the dominant approach
- Not sufficient for full text understanding, but ... IE is not full text understanding
- Single simple pattern matching formalism/engine across all levels of cascade promotes
 - ease of development
 - speed of processing
 - Machine learning of tagging models/patterns ...

Adaptivity via Machine Learning

- A significant problem with IE is creating the
 - Grammar rules
 - Extraction patterns
 - Domain modelsrequired to implement a given extracion scenario
- Creating these manually is time consuming (expensive), error-prone, and requires scarce expertise
- Machine learning (ML) offers the promise of overcoming these problems

Supervised vs Unsupervised Learning

- ML techniques divide into supervised and unsupervised learning techniques
- Supervised learning
 - Training data has associated with it target value/outcome to be learned
 - Advantage: higher levels of accuracy obtainable than with unsupervised techniques
 - Disadvantage: training data must be annotated with target output – may be expensive and time consuming
- Unsupervised learning
 - No need to label training data
 - Advantage: no expensive/time consuming data labelling
 - Disadvantage: lower (sometimes disastrously) accuracy levels

Applying ML to IE

- A variety of ML techniques have been applied to a variety of IE tasks/systems
- Most commonly addressed task is entity extraction, which can be viewed as semantic tagging (a classification task)
 - Usually supervised learning techniques are applied, so training and test corpora are annotated for, e.g. organisation, person, location names
 - Techniques include Hidden Markov Models, Error-driven transformation-based learning, maximum entropy
 - Results match best manually coded rule-based systems
- Increasingly important is the task of learning extraction patterns
- Typically done by shallow parsing + semantic tagging of text then extraction of generalised syntactic/semantic patterns around key scenario verbs
 - Acquired patterns may be filtered manually or automatically

Outline of Talk

- The Text Mining Scenario
- Information Extraction Definition and Scope
- Information Extraction Component Tasks
- Information Extraction Technologies
- Information Extraction Prototype Application
 - Deploying IE
 - TRESTLE System
 - TRESTLE Demo
- Conclusions and Future Directions/Challenges

Deploying IE

- IE is an appropriate technology when:
 - large volumes of text make human analysis infeasible
 - template-oriented information seeking is appropriate (stable information need, narrow domain)
 - conventional IR is inadequate
 - some error is tolerable
 - significance of information need merits cost
- Current levels of error, while no higher than other accepted language technologies (IR, MT), suggest applications which **support** humans, rather than **replace** them, are more appropriate

IE and Information Seeking in Large Enterprises

- To investigate the utility of IE in a real setting have developed an advanced text access facility to support information workers at GlaxoSmithKline
- **TRESTLE** – **T**ext **R**etrieval **E**xtraction and **S**ummarisation
Technology for **L**arge **E**nterprises
- Aim: increase effectiveness of employees in “industry watch” function – current awareness/tracking of
 - People
 - Companies
 - Products – particularly progress of new drugs through clinical trial/regulatory approval process
- Approach: provide enhanced access to *Scrip* the largest circulation pharmaceutical industry newsletter

IE and Information Seeking in Large Enterprises (cont)

- User requirements study at GSK (questionnaire, observation, interviews) revealed 2 key types of information seeking:
 1. Current awareness
 - general updating (what's happened in the industry today/this week)
 - entity or event-based tracking (e.g. what's happened concerning a specific drug or what regulatory decisions have been made)
 2. Retrospective search
 - historical tracking of entities or events of interest (e.g. where has a specific person been reported before, what is the clinical trial history of a particular drug)
 - search for a specific event or a remembered context in which a specific entity played a role

Note: both activities require identification of entities/events in the news = what IE systems do



TRESTLE System Overview

- The system consists of two components

- **Off-line component**

- **LaSIE IE system**

- Input: Scrip texts delivered daily via the Internet
- Output: IE results
 - Named entities: organisation, people, locations, drugs, diseases
 - Scenarios: Person Tracking; Clinical Trials; Regulatory Events

- **Summary Writer**

- Input: Scenario templates
- Output: Single sentence NL summaries of the templates

- **Entity/Scenario Indexer**

- Input: NE annotated texts; Scenario templates
- Output: Indices keyed by NE + date with pointers to source texts

TRESTLE System Overview (cont)

■ On-line component

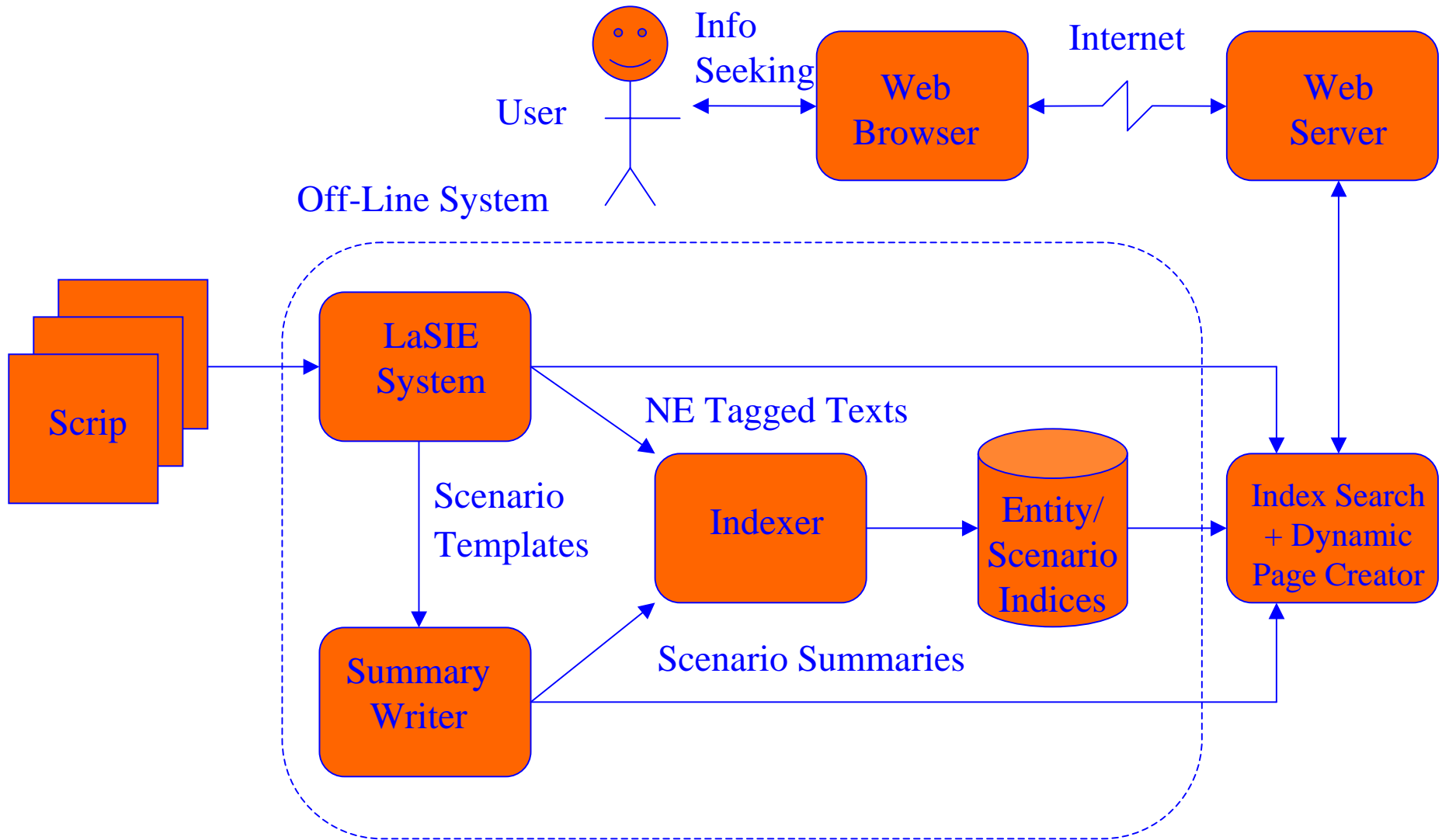
● Browser scripts

- Input: User requests for information
- Output: Results to requests returned from annotated *Scrip* DB

● Entity/Scenario Index Search + Dynamic Page Generator

- Input: User information requests forwarded from Web server + entity/scenario indices + NE annotated texts/summaries
- Output: Relevant HTML pages with link info dynamically generated link information

TRESTLE System Architecture



TRESTLE System Demo



Conclusions and Future Directions/Challenges

- Text mining is a vaguely defined notion which spans a huge range of potential text types, users, information requirements and technologies
- One capability which is a key component in text **content** mining is **information extraction (IE)**
- IE tasks include the extraction of entities, attributes, relations and events into structured representations which are then used for in other applications
- Quantitative evaluation metrics exist for extraction tasks, enabling inter/intra-system comparison and machine learning techniques

Conclusions and Future Directions/Challenges

- Approaches to IE have included
 - “deeper” linguistically motivated techniques
 - “shallower” pattern-matching techniques
- Increasingly machine learning approaches are being used to take advantage of linguistic richness of corpora and to promote adaptivity
- Deploying IE requires acknowledgement of the costs of development and the inevitable errors (“noise”) in the extracted data
 - One way of usefully deploying current IE technology is in “intelligent navigation”, where IE is used to support current awareness or retrospective search in task-oriented intelligence gathering environments

Conclusions and Future Directions/Challenges

■ Future Directions/Challenges include

- Higher accuracy
- Greater speed
- Increased portability across domains/genres
 - extension of ML techniques beyond entity extraction to relation and event extraction
- Support for scenario definition/induction
- Novel applications
- Integration of IE into standard desktop text processing tools

■ End Notes

- Time ... (www.timeml.org)
- E-science + the Grid + the Digital Literature
 - Research support/learning: Ambient text/"terminology tips"
 - Knowledge discovery ...

The End

Further details and papers from: <http://www.dcs.shef.ac.uk/~robertg>